Modeling Operator Performance in Low Task Load Supervisory Domains

by

Armen A. Mkrtchyan

B.S. Electrical Engineering
University of North Dakota, Grand Forks, ND, 2009

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of

Master of Science in Aeronautics and Astronautics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

Author .....
Department of Aeronautics and Astronautics
May 19, 2011

Certified by.....
Mary L. Cummings
Associate Professor of Aeronautics and Astronautics
Thesis Supervisor

Accepted by .....
Eytan H. Modiano
Associate Professor of Aeronautics and Astronautics
Chair, Committee on Graduate Students

Modeling Operator Performance in Low Task Load Supervisory Domains

by

Armen A. Mkrtchyan

Submitted to the Department of Aeronautics and Astronautics
on May 19th, 2011 in partial fulfillment of the requirements for the degree of
Master of Science in Aeronautics and Astronautics.

Abstract

Currently, numerous automated systems need constant monitoring but require little to no operator interaction for prolonged periods, such as unmanned aerial systems, nuclear power plants, and air traffic management systems. This combination can potentially lower operators' workload to dangerously low levels, causing boredom, lack of vigilance, fatigue, and performance decrements. As more systems are automated and placed under human supervision, this problem will become more prevalent in the future. To mitigate the problem through predicting operator performance in low task load supervisory domains, a queuing-based discrete event simulation model has been developed.

To test the validity and robustness of this model, a testbed for single operator decentralized control of unmanned vehicles was utilized, simulating a low workload human supervisory control (HSC) environment. Using this testbed, operators engaged in a four-hour mission to search, track, and destroy simulated targets. Also, a design intervention in the form of cyclical auditory alerts was implemented to help operators sustain directed attention during low task load environments.

The results indicate that the model is able to accurately predict operators' workload. Also, the model predicts operators' performance reasonably well. However, the inability of the model to account for operator error is a limiting factor that lowers model's accuracy. The results also show that the design intervention is not useful for operators who do not have difficulties sustaining attention for prolonged periods. The participants of this study were exceptional performers, since most of them had very high performance scores.

Further research will investigate the possibility of conducting another low task load, long duration study with a more diverse set of participants to assess the impact of the design intervention and to extract personality traits that may affect system performance. Also, the model needs to be revised to take into account operator errors, which can significantly affect performance of HSC systems.

Thesis Supervisor: Mary L. Cummings
Title: Associate Professor of Aeronautics and Astronautics

Acknowledgments

First and foremost, I would like to thank my advisor, Mary Cummings. Thank you for your guidance and support in the past two years. I have learned tremendously from you, and I am grateful for your sincerity and desire to help your students both professionally and personally.

Thank you, Kris Thornburg, for the countless hours you spent reading my thesis. Thank you for your patience and great sense of humor. I could not ask for a better thesis reader. Thank you, Amy D'Agostino, for taking the time to read my thesis and provide feedback.

I am also grateful to the Office of Naval Research for funding my research.

Thank you, Luca, for providing feedback on my research and sharing your thoughts on broader life issues. I look forward to visiting you in Ireland.

Thanks, Yves, for bugging me about my thesis deadline at every opportunity you had and thanks for being such an amazing SCUBA diving instructor. I hope we can dive together again in the near future.

Thanks, Andrew, for not only being a great labmate but also an amazing friend. Thank you for all your encouragement and help with my PhD qualification exams. I greatly appreciate your integrity and willingness to help others and I look forward to a lifetime of friendship.

I am also grateful to my other past and current fellow labmates. Thank you, Jason, Farzan, Jackie, Kim, Fei, Alex, and Christin for your support and friendliness. You have always made me feel comfortable and welcomed.

Thanks to my UROP, Lucy, who helped me with the pilot testing and watching numerous hours of experiment videos.

I would also like to express my gratitude to Sally Chapman, Beth Marois, and Marie Stuppard for all the work they do to make lives of students easier. Also, thank you, Barbara Lechner, for your hospitality during my visit to MIT in the fall of 2008. You were the first person to show me what a phenomenal place the AeroAstro department is.

To the leadership of Sidney-Pacific residence, I want to say a big thank you for being my family away from home. Thank you Michael, Mirna, Amy, Hussam, Matthieu, B, Chelsea, Birendra, William L., Ahmed, George L., Brian for being incredible housemates and exceptional leaders. I hope to get to know each and every one of you better in the future.

I would also like to express my gratitude to Luys Foundation for assistance in the past two years. More specifically, I want to thank Jacqueline Karaaslanian, Anahit, Gayane, and Harout for their hard work to keep Luys scholars happy away from home.

To all my friends in Armenia, you are the people I grew up with and was influenced by. Your pure and unconditional friendship keeps me strong during times of struggle. Thank you, Alik, Karlen, Karo, Levon,

Masis, Mkrtich, Tigran, Zhirayr for being the friends you are even after five years of barely seeing each other.

To my gorgeous Tatev, thanks for your unwavering support in the last six years. You have always had a positive impact on my life. Thank you for always believing in me and making sure I move in the right direction. I look forward to a lifetime of adventures together.

To my younger brothers, Gevorg and Vahagn, you are my best friends and I am honored to be your older brother. Thanks for your love and understanding.

To my parents (in Armenian):

Հայրիկ ու մայրիկ, բառերն իսկապես չեն կարող արտահայտել, թե ինչքան շնորհակալ եմ այն ամենի համար ինչ արել եք իմ համար. Շնորհակալ եմ այն անքուն գիշերների համար, որ անցկացրել եք մտածելով իմ ապագա կրթության մասին. Շնորհակալ եմ, որ զրկել եք ձեզ շատ ու շատ բաներից, որ հասնեմ իմ նպատակներին. Շնորհակալ եմ, որ այդքան հպարտ եք ինձնով. Սիրում ու կարոտում եմ ՁԵԶ.

Last but not least, I want to thank God for all the blessings and opportunities he brings in my life.

Contents

List of Figures

List of Tables

List of Acronyms

ATC	Air Traffic Control
BPS	Boredom Proneness Scale
DES	Discrete Event Simulation
EEG	Electroencephalography
FAST	Fatigue Avoidance Scheduling Tool
FIFO	First-in-first-out
GIGO	Garbage-in-garbage-out
HDS	Hostile Destruction Score
HSC	Human Supervisory Control
LIFO	Last-in-first-out
LOA	Level of Automation
LTL-DES	Low Task Load Discrete Event Simulation
M&S	Modeling and Simulation
MAF	Multidimensional Assessment of Fatigue
MIT	Massachusetts Institute of Technology
MUV-DES	Multiple Unmanned Vehicle Discrete Event Simulation
NPP	Nuclear Power Plant
OPS-USERS	Onboard Planning System for UVs Supporting Expeditionary Reconnaissance and Surveillance
PDF	Probability Distribution Function
SAFE	System for Aircrew Fatigue Evaluation
SAFTE	Sleep, Activity, Fatigue, and Task Effectiveness
SCT	Schedule Comparison Tool
TFS	Target Finding School

UAS-F	Visual Analog Scale for Fatigue
UAV	Unmanned Aerial Vehicle
USV	Unmanned Surface Vehicle
UV	Unmanned Vehicle
V&V	Verification and Validation
VSH	Verran/Snyder – Halpern
WTAI	Wait Time due to Attention Inefficiencies
WUAV	Weaponized Unmanned Aerial Vehicle

1. Introduction

1.1 Motivation

In the past century, the degree of automation in human operated systems has been growing dramatically. Some of the reasons for this growth include increased profitability and productivity, higher reliability and safety, and automation of dangerous and repetitive tasks. Although increased automation is advantageous in many situations, it can also have drawbacks. Ironies and paradoxes of automation refer to the fact that the more advanced the automation is, the more crucial the role of the human operator becomes in successfully monitoring and supervising the automated system (Bainbridge, 1983). Another disadvantage of increased automation is the fear human operators have of losing their jobs (Rifkin, 1995). Even in the early stages of the industrial revolution, a social movement of English textile machine operators, known as Luddites, started destroying automated weaving machines, which were thought to threaten their job security prospects (Spartacus Educational, 2011).

Increased automation can also cause boredom and alertness decrements for operators (Langan-Fox, Sankey, & Canty, 2008). It should come as no surprise that numerous current systems have such a high degree of automation that human operators may have little to do for prolonged periods. Many of these systems can be classified as supervisory control systems, in which "one or more human operators are intermittently programming and continually receiving information from a computer that itself closes an autonomous control loop through artificial effectors and sensors to the controlled process or task environment" (Sheridan, 1992). For a given system, the level of autonomy can vary from manual control to fully automatic. Supervisory control is an intermediate step between the two levels and is shown in Figure 1. There are numerous examples

of highly automated supervisory control systems causing boredom and alertness decrements. One example is the operation of a Predator unmanned aerial vehicle (UAV). In an interview, a Predator pilot said, "Highly skilled, highly trained people can only eat so many peanut M&Ms or Doritos or whatnot…There's the 10 percent when it goes hot, when you need to shoot to take out a high-value target. And there's the 90 percent of the time that's sheer boredom—12 hours sitting on a house trying to stay awake until someone walks out" (Button, 2009). In fact, in a recent study, 92% of Predator pilots reported moderate to total boredom (Thompson, Lopez, Hickey, DaLuz, & Caldwell, 2006). In another example, increased automation in a supervisory control system contributed to low alertness of Northwest Airlines flight 188 crew that in 2009 overflew Minneapolis-St. Paul International Airport by 150 miles (The New York Times, 2009).

Besides the aeronautical domain, there are other areas where boredom and the attentiveness of operators can be problematic in low task load supervisory settings. An example of such a domain is nuclear power plant (NPP) control (Kaku & Trainer, 1992). In describing the everyday

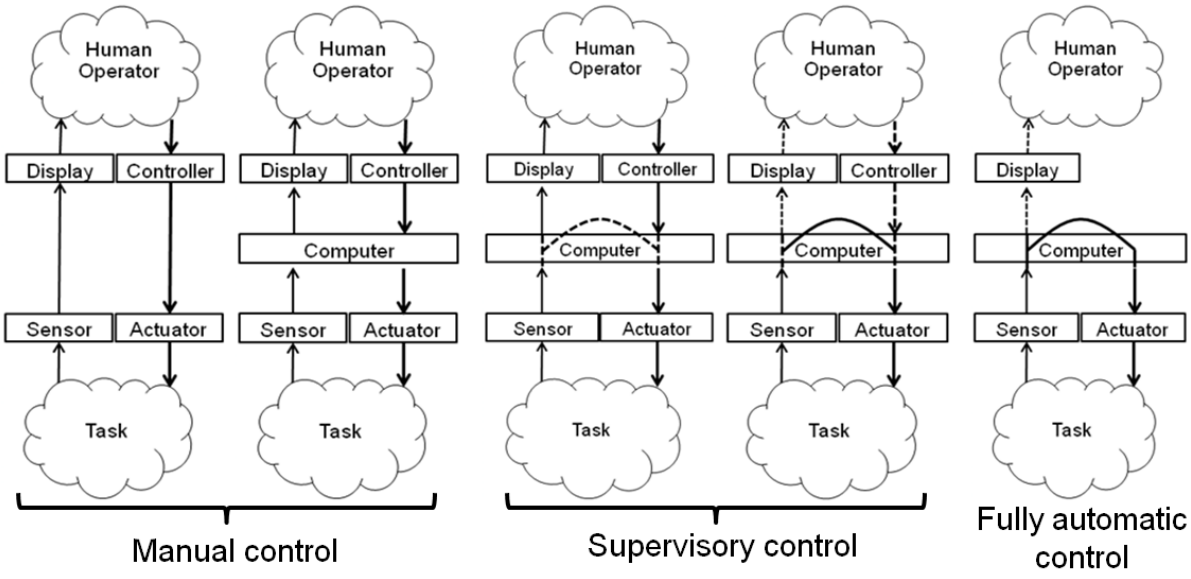


Figure 1: Depiction of different automation levels.

operations of a NPP, an operator said, “People with experience understand how to transition (from low workload to high workload) mostly because you’re waiting for something to go wrong. When the occasion comes people are sometimes excited and thrilled that something happened.” Furthermore, it has been shown that even train engineers (Haga, 1984) and anesthesiologists (Weinger, 1999) experience boredom due to lack of stimulation.

Lack of alertness and boredom, observed in the low task load supervisory domains mentioned above, have further implications. It has been shown that boredom may be a factor that causes complacency, which can be a significant factor that affects performance in supervisory control systems (Prinzel III, DeVries, Freeman, & Mikulka, 2001). Previous studies on air traffic control monitoring tasks showed that participants who reported high boredom were more likely to have slower reaction times and worse performance than participants reporting low boredom (Kass, Vodanovich, Stanny, & Taylor, 2001; Thackray, Powell, Bailey, & Touchstone, 1975). Furthermore, a study of U.S. air traffic controllers showed that a high percentage of system errors due to controller planning judgments or attention lapses occurred under low traffic complexity conditions (Rodgers & Nye, 1993).

Boredom is closely related to vigilance. In fact, it has been shown that participants of vigilance tasks consider these tasks to be boring (Hitchcock, Dember, Warm, Moroney, & See, 1999; Scerbo, 1998). Furthermore, performance decrements due to lowered vigilance have been documented as early as the 1950s (Mackworth, 1950).

To evaluate the effects of low and high task load supervisory control environments on operator and system performance, human-in-the-loop experiments can be conducted. Through these experiments, system designers can evaluate the effectiveness of various design choices.

However, with a large solution space in terms of design options, comprehensive experimentation can be time-consuming and expensive. This cost can arise from both costs of implementing any design changes as well as costs of running the experiments themselves. Moreover, these costs can become prohibitive if the aim is not to test a certain design change, but to search for a design setting that satisfies a certain output condition (i.e., an optimization process). Moreover, human-in-the-loop experimentation can be even more complicated for designs of futuristic systems for which no actual implementation exists (Nehme, 2009). In such cases, experimentation is substituted with approximations from similar systems or must wait for a prototype to first be built.

One alternative to extensive human-in-the-loop experimentation is Modeling and Simulation (M&S). Through M&S, designers are able to predict various trends of system operation, the impact of different variables on system performance, and the effectiveness of design choices. A simplified system engineering diagram (Figure 2) consists of three main stages: (1) requirements definition, (2) system design, and (3) system evaluation. The design stage can be extended to

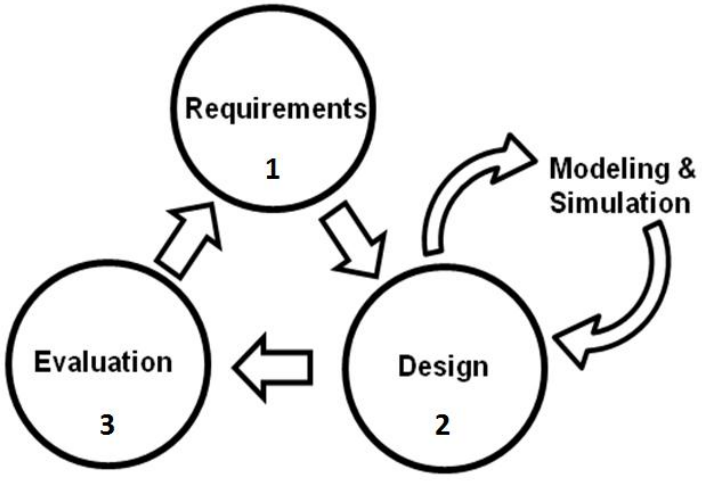


Figure 2: Requirements, Design, and Evaluation loop (Nehme, 2009).

include the M&S loop in order to expedite system's engineering process as well as reduce costs (Nehme, 2009).

One way to extend the design stage to include prescriptive M&S is presented in this research. More specifically, a Discrete Event Simulation (DES) model of a human operator in low task load supervisory domains is presented and discussed as an extension of the design stage. The model is an extension of a previously-developed DES model that is valid under low task load supervisory settings, unlike the previous model, which was not valid under low task load conditions. As operators in supervisory domains interact with automation at discrete set of points in time, DES models are appropriate for modeling operator performance. Moreover, DES models are well-suited to predict operator busyness, number of missed tasks, and delays servicing the tasks (Banks, Carson, Nelson, & Nicol, 2005), which are all important metrics in evaluating performance of operators of human supervisory systems.

1.2 Research Statement

This research was conducted to answer the following research questions:

1. Is it possible to develop a simulation model to successfully replicate and predict operator performance in low task load supervisory domains?

The development of a human performance model for low task load supervisory domains will fill the gap that was left by the previous work (Nehme, 2009). Also, the model could be useful in designing existing systems and evaluating current ones, as described in Section 1.1.

2. Can the model successfully predict the effects of design interventions to mitigate negative consequences of low workload on system performance?

If there are design interventions that can mitigate the negative impact of boredom and vigilance decrement and possibly improve performance of human operators in low task load supervisory systems, it is desirable to predict the effects of these design interventions using the DES model before deciding whether to implement the interventions.

1.3 Thesis Outline

The remainder of this thesis is organized as follows:

Chapter 2, *Background*, discusses human attributes that affect performance in low task load, supervisory domains. More specifically, the impact of boredom, vigilance decrements, and fatigue on performance is presented. Also, results from a previously-conducted boredom study are summarized and the inability of a previously-developed model to predict these results is discussed. Lastly, a new model is proposed to overcome the deficiencies of the previous model.

Chapter 3, *Queuing-Based Low Task Load Discrete Event Simulation Model*, introduces the Low Task Load DES (LTL-DES) model and describes the advantages of the implemented method. Next, various entities of the model are presented, starting with events, service processes, queuing policy, and different attention states. The chapter also discusses the integration of the various entities described above. Finally, the outputs of the DES model are presented.

Chapter 4, *Model Validation and Calibration*, discusses the validation and calibration process of simulation models in general, and the low task load DES model more specifically. A replicative validation process is discussed in detail by using a previously-conducted boredom/vigilance study. Important trends from this study are extracted and a possible intervention to improve system performance in low task load supervisory domains is hypothesized.

Chapter 5, *Low Task Load Experiment and Model Predictions*, describes the experimental setup of the new low task load, long duration experiment, which includes a design intervention to help improve operator performance. The findings of the experiment are presented and discussed, along with comparison to the predictions of the model and results of predictive validation.

Chapter 6, *Conclusions*, describes the motivation for this research, the main findings, accuracy, benefits, and limitations of this research. Also, various applications of the model are discussed. Additionally, information on how this research has furthered knowledge and contributed to tools in modeling human performance in supervisory domains is presented. Finally, future work is discussed.

2. Background

This chapter investigates the effect boredom, vigilance, and fatigue have on human performance in low task load supervisory domains. The results from a previously-conducted experiment that evaluated the impact of low task load settings on human performance are summarized. Finally, the modeling efforts of human performance in supervisory settings are discussed.

2.1 Boredom

As discussed in Chapter 1, boredom is common in low task load supervisory domains. Boredom is defined by Merriam-Webster Dictionary (2011) as a state of being weary and restless through lack of interest. Although boredom can be formally defined, it is not easy to recognize boredom in real life (Svendsen, 2005). Dostoevsky (1997) referred to boredom as bestial and indefinable affliction, while Brodsky (1995) stated that boredom represents pure, undiluted time in all its redundant, monotonous splendor and argued that life to a large extent is boring, since humanity places great emphasis on originality and innovations. Others claimed that they have never been bored (Svendsen, 2005). This variability in opinions shows that boredom can be perceived very differently by different individuals and the effects of boredom can be different as well. More specifically, it has been shown that boredom produces negative effects on morale, quality of work, and performance (Thackray, 1980). However, others have argued that boredom plays an important role in learning and creativity (Belton & Priyadharshini, 2007).

In general, it has been suggested that there are two components of boredom: cognitive and affective (Stager, Hameluck, & Jubis, 1989). The cognitive component represents a human's

perception of the task, i.e., if the task is considered unimportant, the human becomes cognitively disengaged. The affective component represents the human's emotional perception, i.e., feelings of frustration, dissatisfaction, melancholy, and distraction. Not only can boredom cause frustration and dissatisfaction, it can also greatly influence human performance. It has been suggested that a performance decrement can occur because the human withdraws effort and attention resources from a task, even though they are potentially available for performance (Scerbo, 1998).

Boredom has also been classified as either situative or existential (Svendsen, 2005). Situative boredom contains a longing for something that is desired and is often considered to be an emotion, while existential boredom contains a longing for any desire at all and is considered to be a mood (Svendsen, 2005). Also, while situative boredom is usually expressed via yawning, wriggling in one's chair, and stretching out one's arms, profound existential boredom is more or less devoid of expression.

Numerous studies have been conducted to determine the effects of boredom on performance of operators of various systems. One study of air traffic controller (ATC) tasks revealed that under low traffic conditions, the percentage of operator errors due to judgments in planning increased (Rodgers & Nye, 1993). Additionally, ATC operators who reported high levels of boredom had slower reaction times and worse performance compared to operators who reported low levels of boredom (Thackray, Powell, Bailey, & Touchstone, 1975). Moreover, boredom has been shown to adversely affect performance in reading and mathematics tests (Brown & Carroll, 1984). In addition to performance decrements, boredom has also been shown to cause greater anxiety and stress (Colligan & Murphy, 1979; Fisher, 1993), as well as premature death due to cardiovascular disease (Britton & Shipley, 2010; Ebrahim, 2010).

In low task load supervisory domains, the negative effects of boredom can be especially devastating, since the majority of these systems rely on the human operator for decisions that are hard to automate and require human judgment (Brainbridge, 1987). However, generally human operators avoid supervising systems that cause boredom. For example, a boring environment is one of a number of causes that the US Air Force struggles to retain enough UAV pilots (Cummings, 2008). Also, boring environments cause distraction, which further detaches operators from systems they control. For this reason, it is important to identify and measure boredom in supervisory settings. The following section describes several methods of measuring boredom.

2.1.1 Measuring Boredom

To recognize and measure boredom, several methods have been experimentally utilized. In one method, a 3D optical flow tracking system was used to track participants' head positions as they watched a series of boring videos. The participants were rated for boredom events by a group of judges. Ratings and head position data were combined to predict boredom events (Jacobs et al., 2009).

Another study showed the utility of automatically monitoring a student's posture to track the affective states of boredom and high engagement (D'Mello, Chapman, & Graesser, 2007). The results indicated that the affective state of high engagement was manifested through heightened pressure exerted on a seat. Boredom, in turn, was manifested through an increase in the pressure exerted in the back coupled with rapid change in pressure on the seat.

There are also subjective ways of measuring boredom. The Boredom Proneness Scale (Farmer & Sundberg, 1986) is a 28-item survey that measures a person's propensity to being bored. The

authors found strong positive associations with depression, hopelessness, perceived effort, and loneliness. Also, the findings indicated negative associations with life satisfaction.

Other scales measuring boredom include the Boredom Susceptibility Scale (Zuckerman, 1979), Leisure Boredom Scale (Iso-Ahola & Weissinger, 1987), Job Boredom Scale (Grubb, 1975), and Free Time Boredom Scale (Ragheb & Merydith, 2001), among many others.

Furthermore, different personality traits and interests of individuals affect boredom proneness, and hence, performance. One study showed that participants who self-reported high task-related boredom had slower reaction times than participants who reported low task-related boredom (Kass, Vodanovich, Stanny, & Taylor, 2001). In another recent study, subjects were asked to detect flickers on a screen. Subjects who scored low on the boredom proneness scale outperformed people who scored high on the boredom proneness scale and reported less boredom (Sawin & Scerbo, 1995). However, it should be emphasized that boredom is subjective in nature, since it has been shown that mentally demanding situations can cause boredom (Becker, Warm, Dember, & Hancock, 1991; Dittmar, Warm, Dember, & Ricks, 1993; Prinzel III & Freeman, 1997; Sawin & Scerbo, 1994; Sawin & Scerbo, 1995), as can monotonous and repetitive situations. Monotonous and repetitive tasks are also often described as being vigilance tasks, since participants need to maintain continuous alertness to detect the tasks. Vigilance and its measurement methods are presented in the following section.

2.2 Vigilance

Vigilance is defined as “a state of readiness to detect and respond to certain small changes occurring at random time intervals in the environment” (Mackworth, 1957). Vigilance

decrements over long durations were first demonstrated in an experiment in which observers had to monitor movements of a pointer along a circumference of a blank-faced clock (Mackworth, 1950). The monitoring session lasted two hours and the maximal accuracy decrement occurred within just the first 30 minutes. Other studies suggest that not only does the accuracy decrease, but the reaction time of obervers also becomes slower as the time spent on the task increases (Parasuraman & Davies, 1976).

Some researchers stated that vigilance decrements occur under conditions of low workload, when arousal level is low (Manly, Robertson, Galloway, & Hawkins, 1999; Proctor & Zandt, 2008; Struss, Shallice, Alexander, & Picton, 1995). However, a recent study showed that vigilance tasks can be demanding (Warm, Parasuraman, & Matthews, 2008). More specifically, it has been shown that vigilance tasks, rather than being understimulating, are associated with high workload. Furthermore, the vigilance decrement was accompanied by a linear increase in overall workload (Warm, Dember, & Hancock, 1996). To measure the vigilance decrement over time, various techniques were developed, some of which are presented in the follwing section.

2.2.1 Measuring Vigilance

Measuring vigilance can be accomplished by utilizing objective, physiological, and subjective methods (Langan-Fox, Sankey, & Canty, 2009). Each measurment method has its advantages and disadvantages, and one should be careful in choosing the correct method based on the specific situation. Objective ways of measuring vigilance are commonly based on objectively measured performance metrics such as (1) target detection rate, or hit rate, (2) non-target detection rate, or correct rejection rate, (3) failure to detect targets rate, or omission rate, and (4) incorrect identification of non-targets as targets rate, or false alarm rate (Stollery, 2006).

Operator detection times are also used to measure vigilance decrement over time. Several studies on ATC monitoring tasks determined that the time it takes to detect conflict and the frequency of missed conflict increases dramatically over the course of just two hours (Schroeder, Touchstone, Stern, Stoliarov, & Thackray, 1994; Thackray & Touchstone, 1988).

Physiological methods of measuring vigilance include electroencephalographic (EEG) power spectrum changes (Jung, Makeig, Stensmo, & Sejnowski, 1997), cerebral blood flow (Shaw et al., 2009), heart rate fluctuations (Schmidt et al., 2006), galvanic skin response (Chanel, Rebetez, Betrancourt, & Pun, 2008), and assessment techniques such as fMRI (Posner et al., 2008). Although these methods can provide useful data, their accuracy is questionable, since taking into account individual variations and ability to extract signal from noise in the measurements is a challenging problem that has not been fully resolved.

Lastly, it has been shown that several personality dimensions relate to performance efficiency in vigilance tasks (Davies & Parasuraman, 1982). Included in these are introversion-extraversion, field dependence-independence, internal-external locus of control, and the Type-A (coronary-prone) behavior pattern. The findings also indicate that, in general, the performance of introverted observers exceeds that of their extraverted cohorts. Furthermore, field-independent individuals, characterized by good analytical skills and ability to break down a problem into its components, perform better on vigilance tasks than field-dependent observers who are generally less anaytical, not attentive to detail, and think globally. Also, individuals with an internal locus of control (those who view their life as a result of their actions) perform better on a vigilance task than those with an external locus of control (those who believe higher power controls their life). Lastly, performance in vigilance tasks of Type-A individuals who are characterized by a rushed,

competitive, achievement-oriented lifestyle exceeds that of their more relaxed, Type-B counterparts.

Ultimately, vigilance decrement can be caused by either "underload" or "overload" of operators in terms of workload (Pattyn, Neyt, Henderickx, & Soetens, 2008). While underload is associated with boredom and attentional withdrawal, overload is associated with cognitive fatigue and decrease of attentional capacity due to high mental workload. In low task load settings, fatigue is mainly caused by lack of sleep and boredom experienced by operators. The next section presents effects of fatigue on operator performance.

2.3 Fatigue

From a physiological perspective, fatigue is defined as functional organ failure (Berger, McCutcheon, Soust, Walker, & Wilkinson, 1991). However, in supervisory domains, it is more appropriate to talk about psychological fatigue, which is defined as a state of weariness related to reduced motivation (Lee, Hicks, & Nino-Murcia, 1991). Psychological fatigue has been associated with stress and other emotional experiences and may accompany depression and anxiety (Aaronson et al., 1999). Some researchers have integrated the psychological and physiological aspects of fatigue and defined them as the self-recognized state in which an individual experiences an overwhelming sustained sense of exhaustion and decreased capacity for physical and mental work that is not relieved by rest (Carpenito, 1995)

In a study that examined boredom and fatigue experienced by Predator UAV operators, it was found that operators who reported high boredom levels also had high subjective ratings of fatigue. High levels of boredom and fatigue caused slower responsiveness, which resulted in

performance decrements. Also, psychological fatigue is strongly correlated with lack of sleep. Not surprisingly, operators of the morning shift reported the highest fatigue level, since the majority of them expericed sleepiness from early rising (Thompson, Lopez, Hickey, DaLuz, & Caldwell, 2006). The study also found that limiting shift work time to only four hours did not signifiantly influence fatigue and boredom levels experienced by the operators. Several ways of measuing fatigue are presented in the following section.

2.3.1 Measuring Fatigue

Traditionally, measuring fatigue has been hindered because it is a symptom and its subjectivity presents additional measurement difficulties (Aaronson, et al., 1999). In fact, Muscio (1921) was convinced of the uselessness of studies that try measuring fatigue, so he suggested abandoning fatigue measurement studies completely. However, in the last few decades several attempts were made to measure fatigue. Lee et al. (1991) developed the Visual Analog Scale for Fatigue (VAS-F). The VAS-F is an 18-item scale that anchors the measure of fatigue to the current measurement time. It has multiple items to characterize fatigue as it is presently being experienced (e.g., sleepy, fatigued, worn out, energetic, lively).

The Multidimensional Assessment of Fatigue (MAF), developed by Tack (1991), measures subjective fatigue, interference of fatigue with activities of daily living, and subjective distress.

When measuring fatigue, sleep and depression have been identified as correlates (Aaronson, et al., 1999) and can be measured by the Verrad/Snyder-Halpem (VSH) Sleep Scale (Snyder-Halpern & Verran, 1987) and the Profile of Mood States (McNair, Lorr, & Droppleman, 1992), respectively. Other measures of fatigue are McCorkle and Young's Symptom Distress Scale (1978), Rhoten's Fatigue Scale and Fatigue Observation Checklist (1982), Piper's Fatigue Self-

Report Scale (1989), and Fatigue Severity Scale (Krupp, LaRocca, Muir-Nash, & Steinberg, 1989).

To predict operator performance in low task load domains, the various measurement techniques presented above can be utilized to gather data on modeling boredom, vigilance, and fatigue. A model is defined as a representation of a system for the purpose of studying the system (Banks, Carson, Nelson, & Nicol, 2005). Modeling boredom, vigilance, and fatigue is important for having an ability to design systems that take into account the effects these factors can have on operator performance. Previous research on boredom, vigilance, and fatigue modeling is presented in the next section.

2.4 Previous Modeling Efforts

Modeling the effects of boredom, vigilance decrements, and fatigue on operator performance is challenging due to individual behavior differences and the subjective nature of the factors affecting performance. In fact, there are not any well-known models relating boredom and operator performance. However, a vigilance decrement model has been developed by Wellbrink (2004) in which human performance is modeled as a complex adaptive system. More specifically, the model utilizes Multiple Resource Theory (Wickens, 2008) and the human information processing model (Wickens & Hollands, 2000). An experiment conducted with 50 students validated the model by reasonably well predicting observed vigilance decrements. The simple tasks involved in the experiment were (1) Sternberg memory task, (2) cognitive task of computing digits, (3) visual monitoring of a fuel gauge, and (4) clicking an alert button upon

hearing auditory alerts. The biggest limitation of the model was its inability to simulate entirely new human behaviors that are not combinations of observed behaviors.

Fatigue models have been historically more popular and various models have been developed. Most of the models take into account operators' sleep patterns to predict fatigue. One of the most popular models, *Sleep, Activity, Fatigue, and Task Effectiveness* (SAFTE), is presented below, while a more complete list of fatigure models is presented in Appendix A.

The SAFTE model includes sleep reservoir, circadian rhythm, and sleep inertia components that combine additively. The model was developed for use in both military and industrial settings and current users include the US Air Force and Federal Railroad Administration. Also, the SAFTE model has been applied to the construction of a Fatigue Avoidance Scheduling Tool (FAST), which is designed to help optimize the operational management of aviation ground and flight crews (Hursh et al., 2004).

These vigilance and fatigue models were analyzed to evaluate the possibility of using them in predicting operator performance in low task load supervisory domains. Unfortunately, fatigue models are very general, in the sense that these models predict overall operator performance decrement given a sleep schedule and taking into account circadian processes and other factors that are not specific to low task load domains. Also, fatigue models do not take into account specific tasks that operators need to complete, which can vary significantly and can have different effects on system performance.

The agent-based vigilance model (Wellbrink, Zyda, & Hiles, 2004) was developed and validated using very simple memory and cognitive tasks, while the majority of human supervisory

domains include tasks with various difficulty. Hence, this model was not appropriate in low task load supervisory domains.

Another model, which explicitly takes into account various tasks that operators need to complete in supervisory domains was developed to predict the maximum number of heterogeneous unmanned vehicles (UVs) controlled by a single operator (Nehme, 2009). It was based on queuing theory and was implemented as a Discrete Event Simulation (DES) model. Figure 3 shows the representation of this model, called the Multiple Unmanned Vehicle Discrete Event Simulation (MUV-DES). The attributes that MUV-DES captured were grouped by those related to the vehicle team (team structure, role allocation and level of autonomy, and vehicle task allocation), environment, and those related to the human operator (nature of operator interaction). The model contained all major constructs of DES architecture (events, arrival processes for the events, service processes for the events, and queuing policy).

It is important to note that the MUV-DES model can be easily generalized to be used in other human supervisory domains. This can be accomplished by modeling various events of a specific

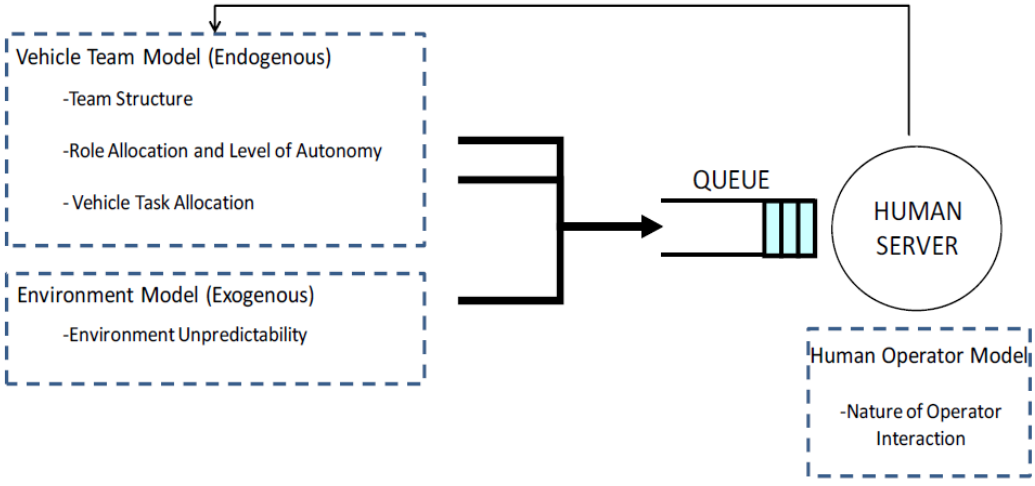


Figure 3: A high level representation of a queuing-based DES model.

domain with respective arrival processes, as well as service processes and other attributes of operator interactions, such as operator inefficiencies. Unfortunately, the model could not be validated in this low task load domain, despite the fact that MUV-DES was validated in medium and high task load situations (Nehme, 2009; Nehme, Kilgore, & Cummings, 2008; Nehme, Mekdeci, Crandall, & Cummings, 2008). Therefore, a new model needed to be developed that was valid in low task load supervisory domains. To gather information for the development of a new model, the previously conducted low task load, long duration study (Hart, 2010) was utilized and is described in the next section

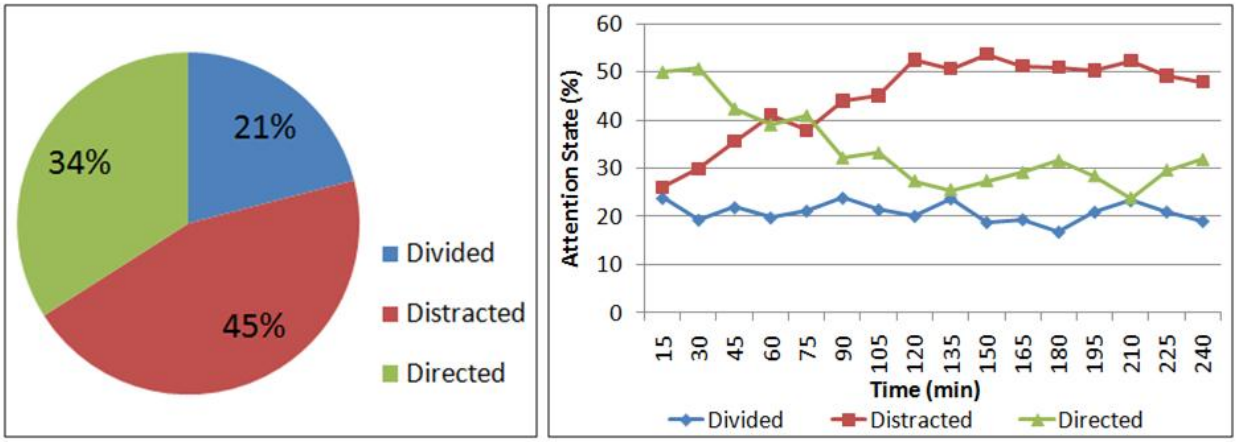
2.5 Previous Long Duration, Low Task Load Experiment

To quantitatively and qualitatively assess the impact of boredom and vigilance decrement on low task load supervisory domains, a long duration experiment was conducted (Hart, 2010). This experiment was conducted using a supervisory interface for controlling multiple simulated unmanned vehicles (Fisher, 2008). In the experiment, a single operator supervised a group of unmanned vehicles in a search, track, and destroy scenario. The operator was assisted by an automated planner for scheduling tasks and planning paths for the UVs. The operator had the ability to edit, cancel, and approve automation generated schedules through a decision support tool. Also, the operator received intelligence information through a chat message window. The study lasted four hours, during which participants' interactions with the interface were logged and participants were video-taped. The key findings of the experiement are discussed in the following section.

2.5.1 Results

One of the most important findings of the experiment deals with the concept of utilization, which is a measure of operator workload. Utilization is defined as percent busy time, where *busy* refers to the time when the operator interacts with the interface. The results of the experiment revealed that the participants interacted with the interface much more than they were required. In fact, their required utilization was about 2% and the average overall utilization was approximately 11%.

Another important finding provided information on operators' attention states. Over the course of the four-hour-long experiment, participants' attention states were approximated by coding their behaviors. Three major attention states were identified: directed, divided, and distracted. The analysis revealed that the majority of the time participants were distracted. Figure 4a shows that on average participants were distracted 45% of the time, directed 34% of the time, and divided only about 21% of the time. Also, over the course of the experiment, participants got more distracted and less directed (Figure 4b). It is interesting to note that the fraction of the divided attention state was almost constant over the the course of the experiment. This means that participants' multitasking abilities did not change significantly in four hours. Also, the study found that the best performers were generally more directed and less distracted than the worst performers. The experiment also provided evidence that distraction does not necessarily degrade performance if managed properly. Particularly, it was observed that one of the participants was able to perform very well by limiting his distraction periods, while on average being distracted about 44% of the time.



a. aggregate attention states **b. attention management over time**

Figure 4: Attention state information of a previously-conducted experiment (Hart, 2010).

These results were used in the development of a DES model designed to predict operator performance in low task load supervisory domains. The next sections highlights the key features of the new model.

2.6 New DES Model

To quantify human performance in low task load domains, a new queuing-based model was developed, taking into account MUV-DES architecture. MUV-DES has been shown to yield accurate results in medium and high task load supervisory domains. Also, queuing models have been successfully utilized in supervisory domains. More specifically, queuing models were used to successfully evaluate pilot's visual behavior when flying a jet airplane (Carbonell, 1966). In other studies, queuing models have been used to evaluate the security and efficiency of air traffic control systems or flight management tasks (Chu & Rouse, 1979; Schmidt, 1978; Walden & Rouse, 1978).

Queuing theory is at the core of DES modeling. A DES simulation is the modeling of a system in which the state variable changes only at discrete points in time (Banks, Carson, Nelson, & Nicol, 2005). DES models are analyzed by numerical methods rather than analytical methods. Analytical methods employ the deductive reasoning to solve the model, whereas numerical methods employ computational methods to solve mathematical models.

The new queuing-based low task load model contains all the components of the MUV-DES model and also additional components to account for the low task load domains. More specifically, the model implicitly takes into account boredom by considering distracted attention states caused by boredom. The model utilizes attention states of operators to account for boredom effects on operator utilization and performance. Lastly, the model takes into account fluctuations in reaction times and task completion times over time, which were observed to vary in the previously-conducted low task load study.

This new model addresses one of the research questions discussed in Chapter 1. More specifically, the model aims to replicate and predict operator performance in low task load supervisory domains. If successful, the model can be utilized to design and evaluate performance of operators in low task load supervisory settings.

2.7 Summary

This chapter introduced boredom, vigilance, and fatigue as significant factors that impact operator performance in low task load supervisory domains and presented ways of measuring each of these factors. Since this research is concerned with developing a predictive model of operator performance in low task load supervisory domains, the effects of boredom, vigilance,

and fatigue cannot be ignored in the development stage of the model. This chapter introduced some of the previously-developed vigilance decrement and fatigue models. However, these models could not be utilized in low task load situations due to their limitations. Another model based on queuing theory was presented, which had been validated in supervisory domains. Unfortunately, the model did not prove to be accurate in low task load domains. To extract trends to develop a new queuing-based model for low task load domains, a previously conducted long duration, low task load study was analyzed. The key findings of the study were discussed, along with their consideration in the new model.

The next chapter presents the new model by thoroughly discussing the different components of the model and various operational modes.

3. Queuing-Based Low Task Load Discrete Event Simulation Model

In this chapter, a queuing-based Low Task Load Discrete Event Simulation (LTL-DES) model is presented. The model can be used to predict the effects that variables, such as task load, level of autonomy, and operator characteristics have on system performance. First, an overview of the LTL-DES model is presented and the necessary assumptions are explained. Next, different constructs of the model are described, which capture different aspects of human supervisory systems. The constructs are then integrated into the model and the operation of the complete model is presented. Lastly, the output metrics of the model are discussed.

3.1 Overview

The LTL-DES captures various attributes of the specific human supervisory control system architecture of interest. These attributes describe both the human operator and the system with which the operator is interacting. On the system side, the level of autonomy (LOA) and required tasks are captured. On the human side, the nature of operator interaction is modeled by taking into account human information processing capabilities, attention allocation strategies, and inefficiencies when performing multiple tasks.

The attributes that describe the human operator and the system are modeled through several DES constructs. A high level representation of the LTL-DES model is shown in Figure 5. The central construct of the model is the *human server* (Figure 5, A), which represents the operator. The human operator is characterized by the *nature of operator interaction* (Figure 5, B), which encompasses the time needed to interact with tasks, as well as operator *attention allocation* (the

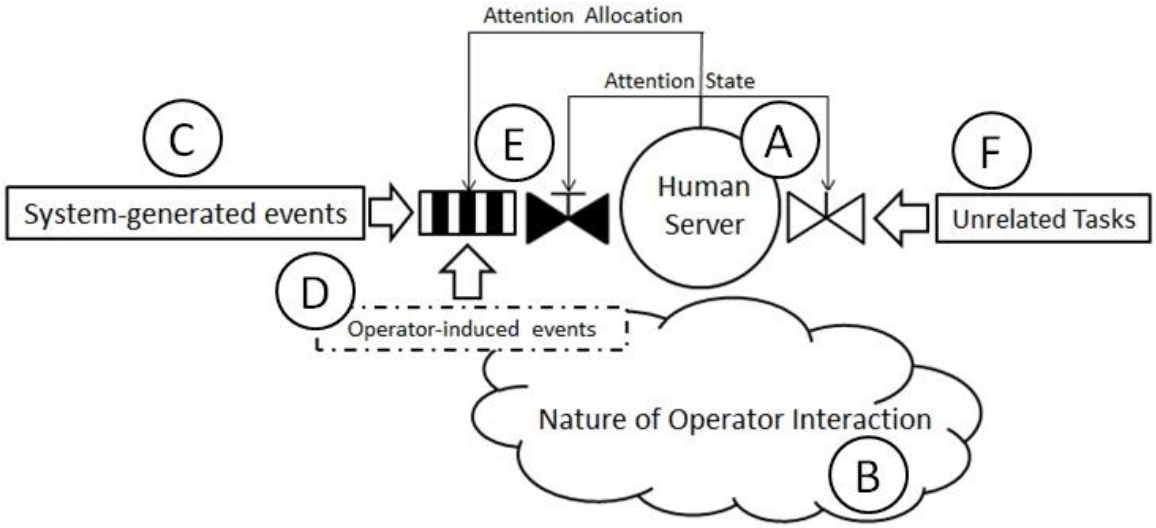


Figure 5: A high level representation of the LTL-DES model.

order with which the operator allocates attention to various tasks). The attention allocation strategy for various tasks and the *attention state* of the operator is governed by the human operator. The model also captures *system-generated events* (Figure 5, C), which are tasks that need to be completed for successful operation of the system. An example of a *system-generated event* is the creation of a task that required by the system. The type and frequency of these events captures the LOA. For example, an automated system with high level of automation will require operator intervention infrequently, while a system with low level of automation will need more frequent operator interaction. Next, the model takes into account *operator-induced events* (Figure 5, D), which are tasks that are not required by the system, but the operator can choose to perform, if desired. Since the automation is not perfect, operator-induced events can improve system performance by targeting weaknesses of the automation. However, operators can also lower system performance by creating additional events that overwhelm the system with tasks that are unnecessary. The system-generated and operator-induced tasks enter the *queue* (Figure 5, E) before they are further processed in the model. Finally, *unrelated tasks* (Figure 5, F) are the

human-generated tasks that do not directly relate to the system and are not specifically modeled, since these tasks lie outside the scope of the control task at hand. In low task load domains, operators resort to these tasks mainly due to boredom that is caused by lack of stimulation and often result in distraction. An example of an unrelated task is the interaction with a personal laptop.

There are several assumptions that the LTL-DES model makes about the human supervisory system. In this case, the operator is assumed to act in a supervisory control nature and interacts with the system at discrete points in time. For this model, only one operator supervises the system, however, it is assumed that the operator can multitask. Also, the system can have multiple sub-systems that are supervised simultaneously.

The LTL-DES contains all major constructs of a DES model, i.e., events, arrival processes for the events, service processes for the events, and queuing policy. An additional construct of the model captures attention states of human operators. This constant is critical in modeling human-system performance, particularly in low task load supervisory domains. In the next sections, these constructs are presented and their usage in the LTL-DES model is detailed.

3.2 Events

An event is an instantaneous occurrence that changes the state of a system (Banks, Carson, Nelson, & Nicol, 2005). Events depend on the type of the system that is being modeled. For example, when modeling a bank teller who is servicing customers, an event can be the arrival of a new customer. When modeling a machine shop operation, an event can be the arrival of a new part order that has been submitted online. In general, events are classified to be either

endogenous or exogenous. The term endogenous is used to characterize events occurring from within a system. Since the LTL-DES was initially developed and utilized to model a futuristic supervisory system in which a single human operator controls multiple unmanned vehicles, examples of different event categories are provided in the context of controlling multiple unmanned vehicles. In this scenario, endogenous events arise expectedly due the nature of the mission and vehicle capability. An example of such an event is replanning vehicle routes after additional tasks have been added. This captures the need for possible rerouting of vehicle paths due to added tasks.

A sub-category of endogenous events is represented by operator-induced events. These events represent the operator's ability to intervene at any point, if desired, to add, delete, or modify existing tasks. Examples of operator induced events are re-planning an automatically generated path to reach a target faster or adding an additional task that is not required by the system. This typically occurs when operators have additional information that is not available to the automation but can be important in achieving mission objectives.

Exogenous events are the environmental events that affect the system. These events arise unexpectedly due to environment unpredictability and create the need for operator interaction, such as an emergent threat area or a meteorological condition, which requires re-planning vehicle trajectories. These are examples of emergent situations that system designers could not account for *a priori*, but are expected given the nature of the UV missions.

Once the events are defined in the model, the rate of their arrival (i.e., how often the events occur) also needs to be defined. The arrival rates of events can be represented by arrival processes, described in the next section.

3.3 Arrival Processes

Each event type in the system has an associated arrival process. The arrival process for an event type is characterized in terms of inter-arrival times of successive events. Arrivals may occur at scheduled times or at random. When the arrival of the event occurs at random times, the inter-arrival times are described by a probability distribution function (PDF). Each event type can have a distinct PDF that describes the arrival process of the specific event type. A typical distribution for random arrivals in supervisory control systems is the Poisson arrival process. The Poisson distribution has been successfully used to model the arrival of events in diverse domains, such as the arrival of phone calls to a call center, arrival of people to restaurants, and the arrival of service orders for product repairs, among many others. It is important to note the memoryless property of the Poisson process. In simple terms, this property means that the current state of a system does not depend on the past states. Mathematically, it can be expressed as:

$$Pr(T > t + \Delta t | T > t) = Pr(T > t). \tag{3.1}$$

In Equation 3.1, T represents the inter-arrival time of events, t is a random time step and Δt is a random period of time. Moreover, the arrival of events in the Poisson process is assumed to be independent. However, in the systems that are being considered in this work, i.e., human supervisory systems, arrival of the events is not always independent and cannot be modeled using the Poisson distribution. Occasionally, the arrival of an event in the LTL-DES depends on the arrival of another event; hence, the memoryless property for probability distributions is not true:

$$Pr(T > t + \Delta t | T > t) \neq Pr(T > t). \tag{3.2}$$

Two arrival processes that are implemented in the LTL-DES, independent and dependent, are described in the next two sections.

3.3.1 Independent Arrivals

The arrival of independent events is not conditioned on the arrival of previously-generated events. For such an arrival, the implementation is straight-forward: a random variable, X_i, represents the inter-arrival time between events of type i. Associated with this random variable, a probability density function of inter-arrival times, $f_{X_i}(x)$, is defined. An example of a type of event that generally can be modeled as having independent arrivals are environmental (exogenous) events, since these events happen in an unpredictable manner and do not depend on events that occurred previously.

3.3.2 Dependent Arrivals

In contrast to independent arrivals, dependent arrival processes are described by the precondition that other events are serviced first. Dependent events can be classified in two categories: (1) events that depend on the same type of event being serviced first and (2) events that are conditioned on other types of events. The first type of dependency between events can be implemented by using the concept of blocking (Balsamo, Persone, & Onvural, 2001). Blocking is used to temporarily stop events from entering the queue before an event of the same type has been serviced. Servicing an event removes the block and the dependent event arrives in the queue. The time between servicing an event and unblocking the dependent event can be either deterministic or stochastic. If it is stochastic, then a probability distribution function, $f_D(d)$, can be used to describe the delay until unblocking. An example of this type of arrival process in the

context of supervising multiple UVs is a replan task that will not be generated unless the previous replan task has been completed.

In the second type of dependency, events of one type are preconditioned on servicing events of another type. This conditioning is implemented by triggering the dependent event after the initial event has been serviced. Triggering the event generates the event of the second type and can be done immediately after servicing the first event, after some fixed time, or can be modeled by using a probability distribution function, $f_T(t)$. It is possible that the triggering of the event can be binary, either occurring or not. In this case, a Bernoulli distribution with probability of p_t can be used to take into account the randomness of triggering the dependent event. An example of a triggered dependent arrival process is the approval of a UV destroying a hostile target, which triggers the arrival of a target destruction confirmation message from the command center supervising the UV mission.

Arrival processes describe the arrival rates of various event types but do not provide information on how long it takes to service these events. Service processes describe the amount of time operators need to interact with different event types, discussed in the next section

3.4 Service Processes

Service processes represent the time that the operator is required to interact with an event. The service times can be uniform or of random nature. In the latter case, they are usually characterized by a probability distribution function. Each event type i can have its associated service distribution, $f_{S_i}(s)$, which captures the variability of a single operator servicing tasks as

well as variability between operators. Occasionally, service distributions are non-stationary and can change over time. For example, when the waiting line in a grocery store is long at peak hours, servicing each customer is usually done faster, resulting in a reduction in service times. To account for the change in service times over time, the parameters describing service distribution can be varied over time. Furthermore, if the variability of service times over time is so large that it cannot be accounted for by using the same service distribution, another service distribution can be utilized to accurately characterize the variation of service times. The majority of systems do not require non-stationary service time distributions to successfully account for the variability of the system However, in the LTL-DES, probability distribution functions are non-stationary to fully account for the observed behavior of human operators.

Servicing an event can have a further impact on the state of the system. First, it can unblock other events and secondly, it can trigger other events with probability of p_t. Not servicing an event can also have a profound impact on the system, since some of the events that are not serviced expire and leave the system without being serviced. This can lower the efficiency and performance of any system and, in some cases, can have devastating consequences. For example, if the event of refueling a UV is not serviced, then the vehicle will eventually crash. It is also possible that an unserviced event can stay in the system and dramatically increase the number of events waiting to be serviced.

Service processes dictate the amount of time an operator needs to complete a task, but do not designate the order in which tasks are serviced. This ordering is identified by the queuing policy, described next.

3.5 Queuing Policy

The queue in the DES models represents event storage. As implemented in the LTL-DES, the size of the queue is unlimited, i.e., the number of events that can be stored in the queue is infinite. The queuing policy defines the order by which the events that are waiting in the queue are serviced. There are various queuing policies that can be implemented in a DES model. Some examples of policies include first-in-first-out (FIFO), last-in-first-out (LIFO), shortest service time first, and highest attribute first, among many others. The FIFO, LIFO, and shortest service time first queuing policies are self-explanatory. The highest attribute first represents a policy in which the high priority events are serviced first (Pinedo, 2002), and priorities of the events can be determined by the designer of the model. A queuing policy that includes a combination of several policies can also be implemented; however, the rules for transitioning from one policy to another must be clearly defined. Since queuing theory is fundamental to the LTL-DES, the queuing notation is presented and the DES model is described using this notation in Appendix B.

In the LTL-DES, besides the queuing policy and service processes that capture the attributes of human interaction with supervisory systems (Figure 5, C), attention states of operators also play a critical role by affecting the flow of events. The next section describes operator attention states more thoroughly.

3.6 Attention States

To properly model the behavior of the operator supervising a highly autonomous system in a low task load domain, the attention states of the operator are taken into account through the LTL-

DES. Modeling operators' attention states is a novel technique utilized in this work. As described in Chapter 2, operators' attention states can be divided into three groups: directed, divided, and distracted. In the representation shown in Figure 3.1, different attention states affect the flow of events from the queue to the human server. Also, depending on the attention state, the attention allocation strategy, which controls the queuing policy, can vary. The operation of the model in the three attention states is presented in the following three sections.

3.6.1 Directed Attention State

In the directed attention state, system-generated (Figure 5, C) and operator-induced events (Figure 5, D) are allowed to reach the human server (Figure 5, A). Simultaneously, on the opposite side of the human server, unrelated tasks (Figure 5, F) are blocked. Hence, the human operator is not paying attention to unrelated tasks and is only servicing and monitoring tasks directly related to the operation of the system. Also, in this attention state, a FIFO queuing policy is implemented. This is a good approximation of the order in which operators service tasks in the directed attention state, since in low task load conditions, the number of events is low and, if directed, operators usually service the events as they see them arriving. In low task load control environments, it is rare that more than one event arrives in a very short time frame. However, if few events arrive in a short time period, operators usually service these tasks in the order they arrive. It is important to note that the LTL-DES model is flexible and allows the queuing policy to be changed, if for some reason it is assumed that the operator has another strategy for servicing events.

3.6.2 Distracted Attention State

In the distracted attention state, the human operator does not pay attention to the system; therefore no system-related events reach the human server (Figure 5, E), resulting in all system-generated events (Figure 5, A) accumulating in the queue. Some of the events expire and leave the queue if the wait time is longer than the time available to complete the task. Also, since the operator does not pay attention to the interface, operator-induced events are not generated (Figure 5, B). Finally, when the operator switches from a distracted to a directed attention state, the queuing policy changes from FIFO to a priority-based queuing policy. The priority-based policy stays in effect until all the events that were in the queue at the time of switching are serviced. This change in the queuing policy captures the fact that after spending some time in the distracted attention state, events accumulate in the queue. It is assumed that when switching to a directed attention state, operators service the accumulated events according to a priority-based queuing policy, which is defined based on mission specifications. Once all the accumulated events are serviced, the queuing policy switches back to FIFO policy, if the operator remains in a directed state.

3.6.3 Divided Attention State

In the divided attention state, it is assumed that the operator is multitasking. It has been observed that most operators in this state do not create any operator-induced events. However, the operators service system-generated events, although events wait longer in the queue to be serviced due to operator attention inefficiencies, discussed in the next section.

3.6.3.1 Wait Times Due to Operator Attention Inefficiencies

The concept of wait times due to attention inefficiencies (WTAI) has been previously used to model the performance of UV operators (Donmez, Nehme, & Cummings, 2010; Nehme, Crandall, & Cummings, 2008). WTAI represents the effects of low situation awareness on task wait times (Nehme, Crandall, & Cummings, 2008). Previously, it has been used in conjunction with the busyness level of operators to account for additional delay in servicing events (Nehme, 2009). More specifically, it was assumed that the delay is the greatest when the operator is either very busy or is almost idle. The wait time is the shortest when the operator is moderately busy. The concept of linking the busyness level of operators and WTAI proved to be critical in successfully modeling human operators controlling multiple unmanned vehicles (Nehme, 2009).

However, in low task load domains, the busyness level is low to start with and does not vary significantly; hence, conditioning WTAI on the busyness level is not viable. Fortunately, it has been observed from the experiment described in Chapter 2 that in low task load domains, WTAI is related to the time that the operator spends supervising a system. Specifically, in that low task load experiment (Hart, 2010), the wait time of events was the shortest in the beginning and in the end of the experiment. Based on this observation, it was assumed that in the divided attention state, WTAI can be represented as an inverted U-shaped function that is related to the time of experiment. Figure 6 demonstrates this relationship, which is modeled by using a parabolic function, where T is the duration of the experiment and W_1 and W_2 are the minimum and maximum durations of wait times, respectively. It is also assumed that WTAI is zero in the directed attention state, since the operator is engaged in monitoring the system and attention inefficiencies are negligible. Lastly, in a distracted attention state, events are not serviced; hence, the wait time of events grows in accordance to the time spent in the distracted attention state.

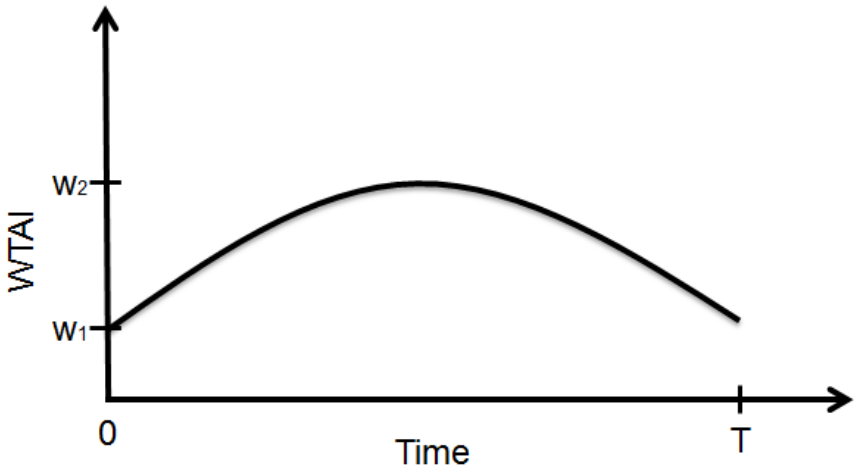


Figure 6: WTAI – time relationship (applies to divided attention state).

An important step in developing a valid model is to successfully integrate the various components described above. The next section presents the integration process with emphasis on the low task load experiment discussed in the previous chapter.

3.7 Integrating Model Components

In the LTL-DES, different constructs of the model are integrated in such a way so that the model captures different attributes of human supervisory systems described in Section 3.1. To more thoroughly show the process of modeling the various attributes through the constructs discussed earlier, the model is discussed in the context of supervising multiple unmanned vehicles. The low task load study described in Chapter 2 provides the necessary information for estimating parameters of the LTL-DES model.

First, the events in this model capture the different tasks that are available for the human to perform, within the study described in Chapter 2. The endogenous events are:

- Creating/editing search tasks – Operators need to create/edit search tasks to prompt the vehicles to search a specific area to find friendly, unknown or hostile targets.
- Replanning – Operators can replan to assign vehicles to various search tasks.
- Read/Respond to chat messages – Operators need to respond to chat messages that are sent by the command center.

Exogenous events are:

- Target identification – Operators need to identify targets that vehicles find.
- Weapons launch approval – Operators need to approve the destruction of hostile targets.

Furthermore, events that are time critical and need to be serviced within a predetermined amount of time expire and leave the queue. In the LTL-DES, the following events may expire:

- Chat message events expire after a predetermined amount of time if not serviced.
- Replan events prompted by the system expire once the next replan event is prompted.

It should also be noted that operators can induce more events by creating/editing search tasks and replanning. As described earlier, operator-induced events are taken into account in the directed attention state. In the divided attention state, most of the time operators only serviced system-generated events and only rarely created additional events. For this reason, the model assumes that additional events are not created in the divided state of its operation. The different types of events have associated probability distribution functions that characterize the inter-arrival time of the events (Appendix C). Furthermore, dependencies among the events are modeled in the LTL-DES. More specifically, the arrival of each type of endogenous event is conditioned on the same type of event being serviced first. For example, if the operator has not serviced a replanning event that is in the queue, another replanning event will not arrive to the queue until the first

replanning event has been serviced. Also, servicing some of the events triggers the arrival of other events. For example, servicing a weapons launch approval event triggers the arrival of a chat message event, since a confirmation message is sent once the missile destroys a target.

To model the service times, probability distribution functions are used to describe operator interaction with the events. Each type of event has an associated probability distribution function, which was computed by using the observed data from the experiment discussed in Chapter 2. It was also observed in the experiment that over time, event service times became shorter; however, operators' reaction times were the fastest in the beginning and in the end of the experiment and slowest in the middle. To account for change in service times, the probability distribution functions characterizing service times are non-stationary and change every hour in the LTL-DES, based on the observed service times (Appendix C).

The flow of events is greatly affected by the attention states. As described earlier, in the directed attention state, all the possible events, except the unrelated tasks, are generated and the model operates at its maximum capacity. In the divided attention state, only system-generated events are taken into account. Also, the servicing of tasks may be delayed due to WTAI. Finally, in the distracted attention state, the flow of events stops in the queue while the system continues generating endogenous and exogenous events. If the operator spends considerable amount of time in the distracted attention state, then the performance of the system can be expected to decline. To measure the performance of the system, performance metrics were defined. These metrics are the outputs of the model and are presented in the next section.

3.8 Model Outputs

In general, primary steady-state measures of a queuing-based DES model are the average number of events in the system and in the queue, the average time events spend in the system and in the queue, and the server utilization for the whole duration of the mission and for shorter time intervals. In the LTL-DES model, from the steady-state measures mentioned above, utilization is used as a measure of operator workload. It is calculated as the ratio of the time the server is busy servicing tasks divided by the total duration of the simulation. For a single server queuing system, such as the LTL-DES, the long run server utilization (ρ) is equal to the average event arrival rate (λ) divided by the average service rate (μ).

$$\rho = \frac{\lambda}{\mu} \tag{3.3}$$

For the queuing system to be stable, the arrival rate must be less than the service rate, i.e., $\lambda < \mu$. If the arrival rate is greater than the service rate, then $\rho = 1$. In real-world situations, this can happen when operators are asked to do more than they can handle. One way to alleviate the saturation of servers is to increase the number of servers (i.e., operators).

In the LTL-DES model, average event wait time in the queue is also calculated. To find the average time events spend in the system, we define $W_1^Q, W_2^Q, \dots, W_N^Q$ to be the time each event spends in the system, where N is the number of arrivals during $[0, T]$. Hence, the average time spent in the queue per event will be:

$$\widehat{w}_Q = \frac{1}{N}\sum_{i=1}^{N} W_i^Q \tag{3.4}$$

As $N \to \infty, \widehat{w}_Q \to w_Q$, where w_Q is the steady-state time spent in the queue. For stable queuing systems, w_Q must be bounded, otherwise wait times will grow indefinitely. It has been shown that average event wait time and operator utilization can allow for comparison across multiple applications (Pina, Cummings, Crandall, & Della Penna, 2008).

Besides the basic metrics that can be captured by DES models, system designers can also create mission specific metrics to evaluate the impact of variables of interest. In the LTL-DES model, the combination of directed and divided attention states, as well as average task wait time was utilized to assess operator performance. More specifically, the performance score was computed as shown below.

$$PS = 2.371P_{Dir} + 2.318P_{Div} - 0.094w_Q \tag{3.5}$$

In Equation 3.5, PS represents the performance score, P_{Dir} and P_{Div} represent the percentage of time spent in directed and divided attention states, respectively. The relationship between directed, divided attention states and performance score was extracted from the previously-conducted low task load, long duration experiment (Hart, 2010).

3.9 Summary

The LTL-DES provides a novel way of modeling system performance of low task load supervisory domains. The model has all the basic constructs of a traditional DES model, i.e., events, arrival processes, service processes, and various queuing policies. Through these constructs, the supervisory system and the human operator are described. In addition to the basic constructs, a new modeling approach is used in which the LTL-DES takes into account the

attention states of operators, which are used to temporarily stop the flow of events and change the queuing policy. Another factor that differentiates the LTL-DES model from a previously-developed DES model (Nehme, 2009) is the use of non-stationary service time distributions to more accurately model human operators. Lastly, the model utilizes a novel approach in modeling wait times due to attention inefficiencies by conditioning these wait times on the total mission time, rather than on predetermined intervals during the mission (Nehme, 2009). The model outputs various performance metrics that can be used to evaluate the system. More specifically, the number of events serviced, utilization, and average event wait time in the queue, as well as a mission-specific metric are utilized in the LTL-DES model.

To confirm that the model represents true system behavior and to increase the credibility of the model, it needs to be validated. The validation of the LTL-DES model was conducted by using a historical data set and a human-in-the-loop experiment. This process is presented in the next chapter.

4. Model Verification and Validation

One of the most important and difficult tasks in the development of a simulation model is the verification and validation (V&V) process (Banks, Carson, Nelson, & Nicol, 2005). If the V&V process has not been performed or the model failed to satisfy V&V requirements, designers should be skeptical of using the model to make design recommendation or judgments about the operation of the system. As defined by Banks et. al. (2005), verification is concerned with building the model correctly. The goal of the verification process is to guarantee that the conceptual model matches the computer representation, i.e., the simulation software is implemented correctly. In contrast, the validation process is concerned with building the correct model. The objective of validation is to confirm that the model accurately represents the real world system. It is often too costly and time consuming to determine that a model is absolutely valid over the complete domain of its intended applicability. Instead, tests and evaluations are conducted until sufficient confidence is obtained that a model can be considered valid for its intended application (Sargent, 2007).

It is also important to understand that the outcome of V&V should not be considered as a binary variable, i.e., the model's accuracy is neither perfect nor completely imperfect (Balci, 2003). Modeling and Simulation (M&S), by definition, is an approximation of the real system; therefore, it is only logical to assume that the results of the model can either describe the real-world system with sufficient accuracy or not.

Lastly, one should take into account the possibility that the V&V process for all the sub-systems of a model yields sufficient results, but the model as a whole fails to meet the V&V criteria

(Balci, 1997). Ultimately, the most important decision criterion when determining the accuracy of a model is to determine the accuracy of the whole.

To describe the V&V process of the LTL-DES, the next two sections describe both the verification and validation in general and in the context of the LTL-DES.

4.1 Verification

To assure that the conceptual model (i.e., system structure, system components, parameter values, simplifications, etc.) of the system is reflected sufficiently in the operational model, several techniques are suggested. Some of the techniques are presented below and have been utilized in the verification process of the LTL-DES.

- The assumptions of the operational model should be checked by a subject matter expert who is not the developer of the model (Banks, Carson, Nelson, & Nicol, 2005).

The LTL-DES assumptions, presented in Section 3.1, were reviewed by at least two researchers who had in-depth knowledge of queuing-based models and UV control systems. Their feedback confirmed the reasonableness of the model assumptions.

- Graphical interfaces representing the operational model are recommended and can simplify the task of understanding the system (Bortscheller & Saulnier, 1992)

Animations of the LTL-DES in three different attention states were created that showed the flow of events in the model. These animations proved to be very useful in understanding and visualizing the operation of the model. A graphical representation of the divided attention

state, as applied in the LTL-DES is shown in Figure 7. The figure shows the human server attending to both system-generated and unrelated tasks. Also, self-imposed events are not added to the queue, which contains two different events, represented by the diamond shaped figures.

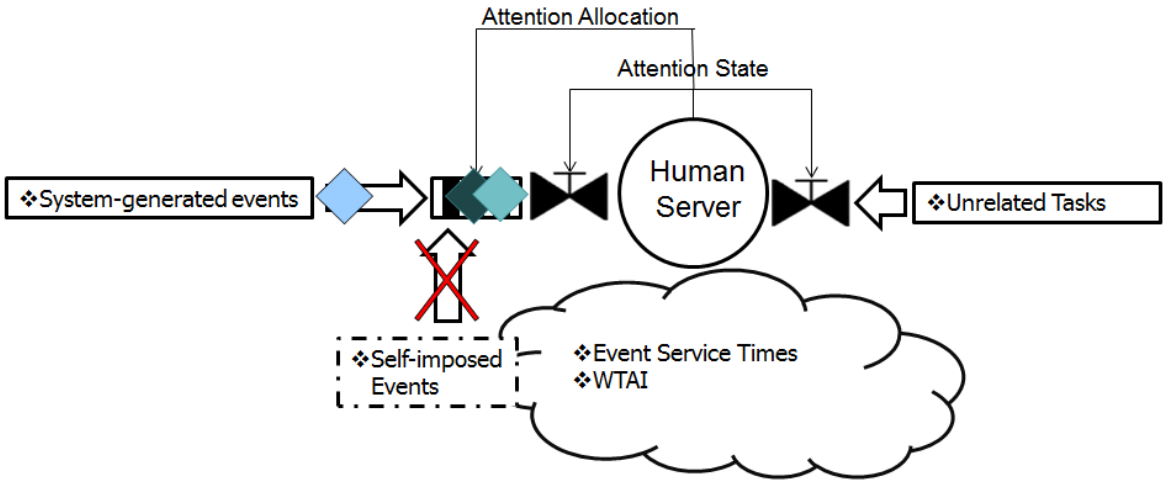


Figure 7: Graphical representation of the divided attention state.

- The outputs should be examined for reasonableness.

During various stages of the LTL-DES development, some of the outputs, such as utilization and number of events serviced, were monitored for unusual values. For example, the number of events were monitored during the development process of the model. If the number was significantly different from an expected value, then there was a high probability the simulation was not implemented correctly and the cause of the discrepancy needed to be identified. This process helped ensure that the operational model reasonably reflected the expected results of the conceptual model.

These techniques are all subjective model verification methods. However, for some models, there are objective ways of verification. For example, when verifying queuing-based systems with Markovian arrival and service processes, several performance measures can be calculated analytically and compared to the simulation results. Unfortunately, the LTL-DES is too complex to use analytical techniques to compute steady-state measures of performance, hence, only the above mentioned techniques were utilized to successfully verify the model.

Once the verification process was completed, the model needed to be validated. In general, two types of validation exist: replication and prediction validation (Balci, 2003). The replication validation of the LTL-DES is discussed in the next sections, whereas the prediction validation is presented in Chapter 5.

4.2 Replication Validation

Model validation goes hand in hand with model calibration, which is the iterative process of comparing the model to the real system and making adjustments, if necessary (Banks, Carson, Nelson, & Nicol, 2005). Figure 8 graphically depicts the validation and calibration process. As one might expect, validation and calibration can continue indefinitely by continuously recalibrating the model to characterize new systems. Ultimately, the modeler's judgment, time, and the cost associated with recalibration dictates the number of readjustments needed to have a sufficiently well-calibrated model.

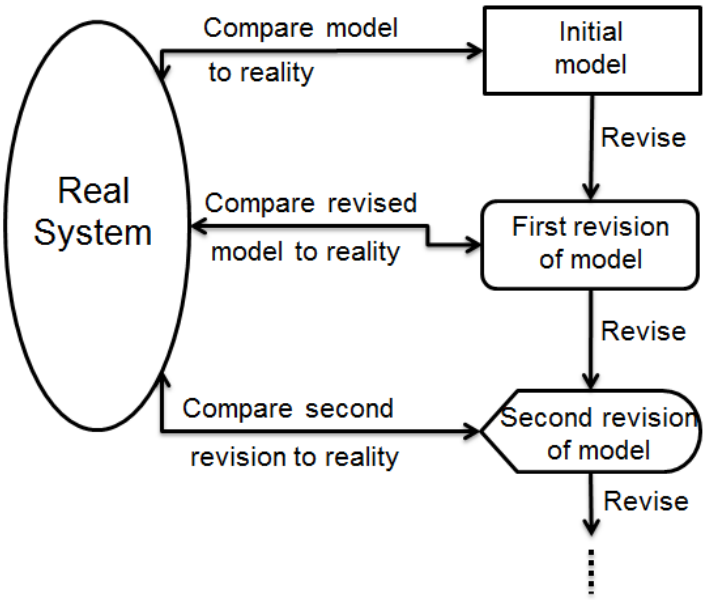


Figure 8: Model validation and calibration (Banks et al., 2009).

To simplify the validation and calibration process, a three-step approach has been proposed by Naylor and Finger (1967). The three steps are:

- The model needs to have high face validity.
- Assumptions of the model should be validated.
- The input-output transformations of the model should be compared to input-output transformations of the real system.

These steps are discussed in detail in the next sub-sections.

4.2.1 Face Validity

The model is considered to have high face validity if, on the surface, it appears to model the system it is supposed to be modeling (Proctor & Zandt, 2008). For the LTL-DES, high face validity was established by comparing the actual and expected responses of the model when varying input variables to the model. For example, the percentage of directed attention state was varied and predicted utilization was examined against the expected response, i.e., on average

more directed operators should have higher utilization compared to less directed operators. Also, subject matter experts in the area of modeling human supervisory control domains have indicated that the model has all the necessary components to successfully model human interaction with autonomous systems in low task load supervisory domains. Lastly, the iterative calibration process of the LTL-DES model provided information about the strengths and shortcomings of the model. The biggest shortcoming of the earlier versions of the model (Nehme, 2009) was the inability to account for operators' low task loading, which is crucial in predicting performance in low task load domains. The model was later modified to account for specific attention states. Once the face validity of the LTL-DES had been established, the assumptions of the LTL-DES were validated, discussed next.

4.2.2 Validation of Model Assumptions

The assumptions made while developing the LTL-DES model can be grouped in two categories: structural and data assumptions. Structural assumptions refer to the operation of different model components and simplifications made when integrating the components. The structural assumptions are validated by comparing observations to the operation of the model. In the case of the LTL-DES, observations were acquired from the data set generated by the long duration, low task load experiment described in Chapter 2 (Hart, 2010). The following are the main structural assumptions:

- The queuing policies implemented in the DES model were compared to the observed strategies that operators utilized to service tasks. The results of the comparison validated the assumptions that FIFO and priority-based queuing policies are sufficient for modeling operators' task selecting behavior.

- The structural assumptions made when implementing the three attention states (Section 3.7) were also validated by examining operators' observed behavior. More specifically, it was observed that in the directed attention state, operators serviced both system-generated events and induced additional events not required by the system. In the divided attention state, only system-generated events were serviced. Lastly, in the distracted attention state, events were not serviced at all.
- The assumption that human operators act as serial processors of tasks was validated based on the cognitive nature of the tasks. More specifically, higher level (cognitive) tasks are attended serially, whereas more low level, perception-based tasks can be processed in parallel (Liu, Feyen, & Tsimhoni, 2006).

Validating the data assumptions was conducted by computing inputs of the model using the historical data set. Input modeling is one of the most critical stages in model validation, since inaccurate model inputs invariably yield invalid results, even when the structure of the model is accurate. This concept is also known as garbage-in-garbage-out (GIGO). For the LTL-DES, inputs to the model are inter-arrival and service distributions for various tasks, along with distributions of the three attention states. In the past, probability distributions were calculated by (1) hypothesizing an appropriate distribution, (2) estimating the parameters for the hypothesized distribution, and (3) conducting goodness-of-fit tests to validate the assumed statistical distribution (Banks, Carson, Nelson, & Nicol, 2005). Currently, several statistical software packages can quickly and accurately identify the statistical distribution that best fits collected data (e.g., MATLAB©, EasyFit©). To model the input data for the LTL-DES, the statistical software package EasyFit© was used to identify the distributions that best fit the observed data of the previous low task load study. The Kolmogorov-Smirnov goodness-of-fit test was utilized to

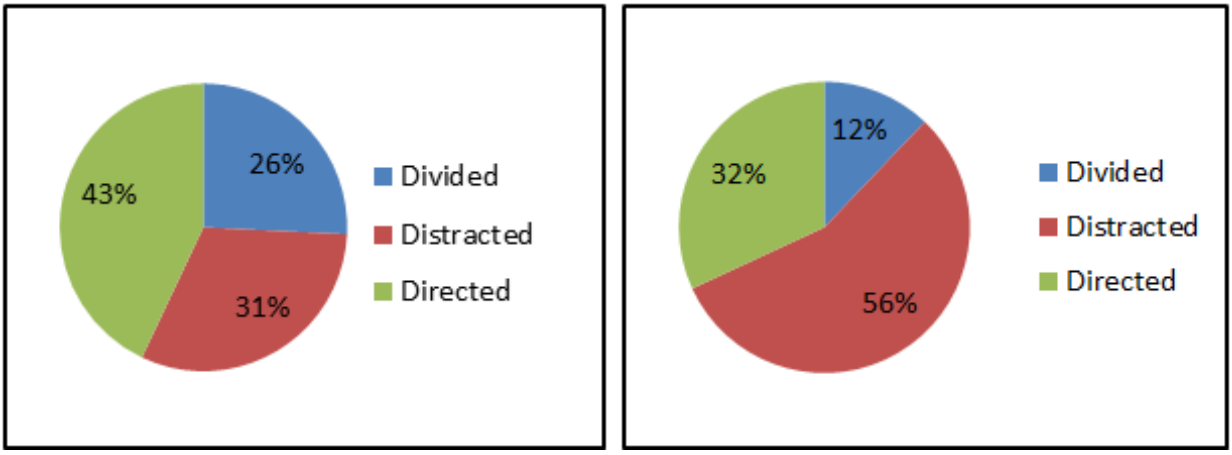
assess the suitability of a chosen distribution. A significance level of 0.05 was selected to either reject or accept the goodness-of-fit test. The parameters of the probability distribution functions selected as inputs to the model for the service times and inter-arrival times are valid estimates of the observed data and are presented in Appendix C.

Once the inputs of the model were selected, the next step to validate the LTL-DES was to evaluate the input-output transformations.

4.2.3 Validating Input-Output Transformations

The input-output transformation is the most important stage in validating a simulation model and can be considered as the only objective way of testing the model. In this stage, the model accepts values of input parameters (distributions) and transforms them to performance measures (output variables). The transformations of the model are then compared to the actual observations to evaluate the validity of the model. The LTL-DES was validated using an historical data set collected from the study presented in Chapter 2 (Hart, 2010). The input parameters of the model were generated based on this data set, as explained in the previous section and are discussed more thoroughly in the next chapter.

As described in Chapter 2, the best performers in the study (whose performance scores were one standard deviation above the mean) were generally more directed than worst performers (whose performance scores were at least one standard deviation below the mean). Moreover, the difference between the best and worst performers' attention allocation strategies was so profound that two different versions of the LTL-DES were developed to account for significant differences in attention allocation strategies. Figure 9 shows why it is important to model the best and



a. best performers **b. worst performers**

Figure 9: Attentions states of (a) best and (b) worst performers.

worst performers separately. It is easy to see that the best performers were better multitaskers and more directed than the worst performers. Also, the best performers were more consistent in their attention switching behavior, i.e., the amount of time they spent in a directed attention state was correlated with the number of times they switched to a directed attention state, while there was no correlation for the worst performers. Hence, the two versions of the model took into account the overall percentages of attention states, as well as the switching pattern between the states. Also, having the ability to predict performance metrics of the best and worst performers provides important information about the range of possible outcomes. It is also essential to know the behavior of an "average person", which is also modeled in the LTL-DES.

After developing the LTL-DES model for the average person and for the best and worst performers, 500 iterations of the LTL-DES were conducted. Each trial simulated one operator involved in a four hour mission supervising multiple UVs. After each simulation trial, initial values of the model parameters were randomized to simulate different human behavior. Once all 500 trials were complete, all the generated data was transferred to a MATLAB© file for analysis. For various performance measures (number of events serviced, utilization, and mission

performance score) the observed values were compared to the model outputs to establish model validity. The outcomes of these comparisons are presented in the next sub-sections.

4.2.3.1 Number of Events

The number of events that operators service during the experiment represents a basic metric that allows designers to check the validity of the model. A significant difference between the number of observed and model predicted events indicates that most likely the inter-arrival distributions were not constructed correctly. Another possibility for the discrepancy can be inaccurate modeling of the attention states, since these states affect the flow of events into the human server (Section 3.7). The predicted and observed number of serviced events for all participants, as well as the best and worst performers are shown in Figures 10, 11, and 12, respectively. In the case of the best performers, the observed mean of the number of serviced events was 442.6 (s.d 50.86). The model predicted the number of events for the best performers to be 455.9 (s.d. 17.4), indicating that the model predictions fall within the standard deviation of the observed number of events. There was no statistical difference between the means ($t(4) = -0.29, p = 0.79$). For the worst performers, the observed mean of the number of events was 352.4, with a very large standard deviation (141.3). The predicted mean value was 335.3 (s.d 25.6) and there was no statistical difference between the means ($t(4) = 0.05, p = 0.97$). For the overall average, the observed number of events was 404.7 (s.d. 138.3). The predicted number of events was 391.62 (s.d. 28.5). Although there was no statistical difference between the predicted and observed number of events ($t(24) = -0.52, p = 0.61$), the model is unable to capture the large variability of the observed number of events. This is mainly due to a great variability of different strategies that operators employ to supervise highly autonomous systems.

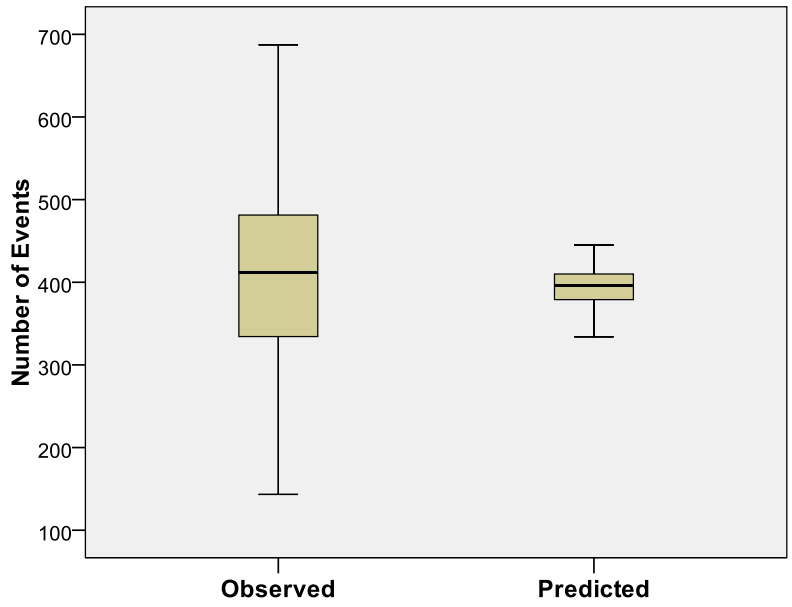


Figure 10: Observed and predicted number of events for all.

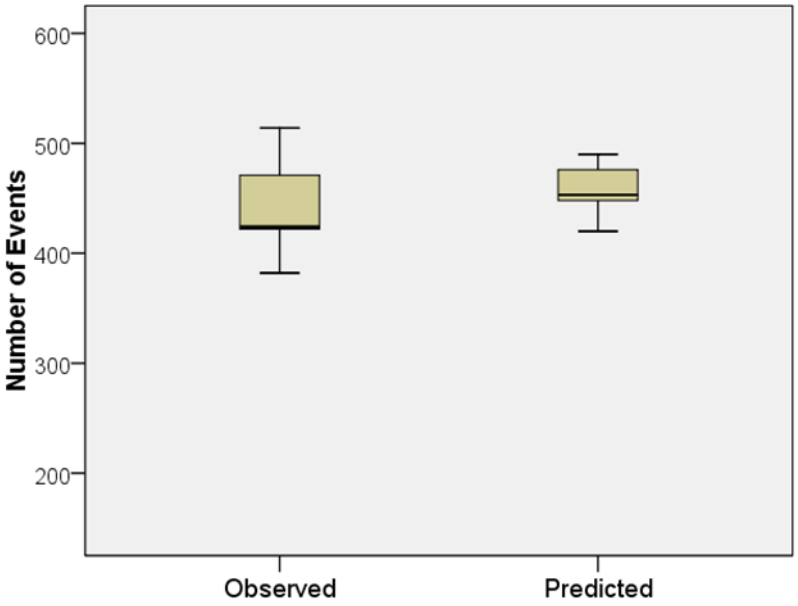


Figure 11: Observed and predicted number of events for best performers.

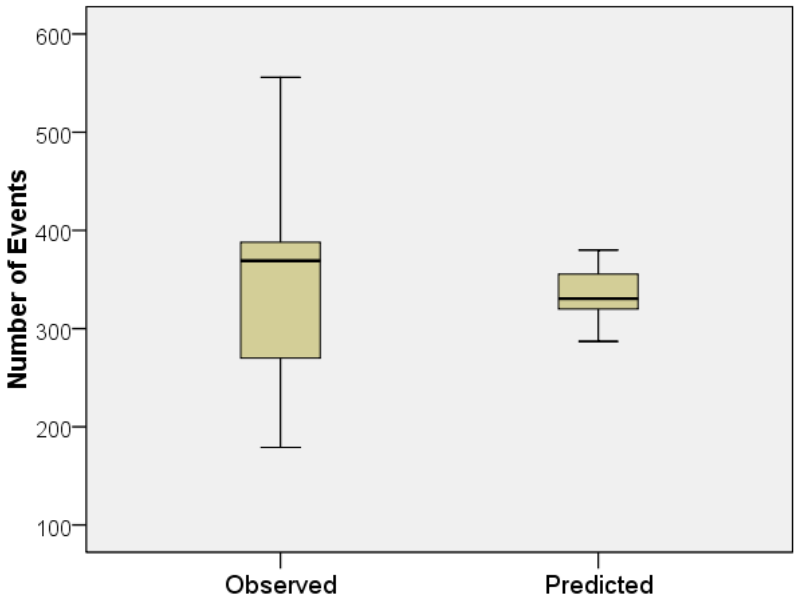


Figure 12: Observed and predicted number of events for worst performers.

4.2.3.2 Utilization

Utilization can be used to approximate operators' workload. Therefore, it is crucial for the LTL-DES to accurately replicate observed values. Figures 13, 14, and 15 depict the observed and

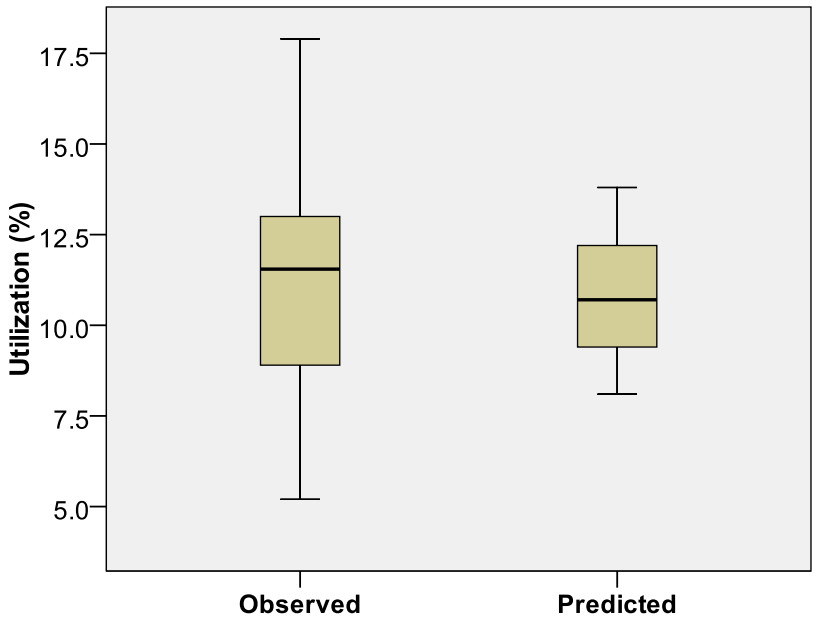


Figure 13: Observed and predicted utilization for all.

predicted utilization for all, best and worst performers, respectively.

For overall average, the observed mean utilization was 11.4% (s.d. 3.36%). The predicted utilization was 10.9% (s.d. 1.6%) and there was no statistical difference between the means ($t(24) = 0.81, p = 0.43$). For the best performers, the observed average utilization was 13.1% (s.d. 3.8%), whereas the predicted average utilization was 12.6% (s.d. 0.7%). For the worst performers, the observed average utilization was 11.22% (s.d. 4.1%), whereas the predicted average utilization was 9.82% (s.d. 1.4%). In both cases, the model predicts the utilization sufficiently well, since model predictions fall within one standard deviation of the observed values and no statistical difference was observed (best performers: $t(4) = 0.25, p = 0.83$,worst performers: $t(4) = -0.73, p = 0.51$). Nonetheless, the predicted standard deviation is about 3-4 times smaller than the observed standard deviation. This is again due to the very large variability of different strategies that operators employ during human-in-the-loop experimentation.

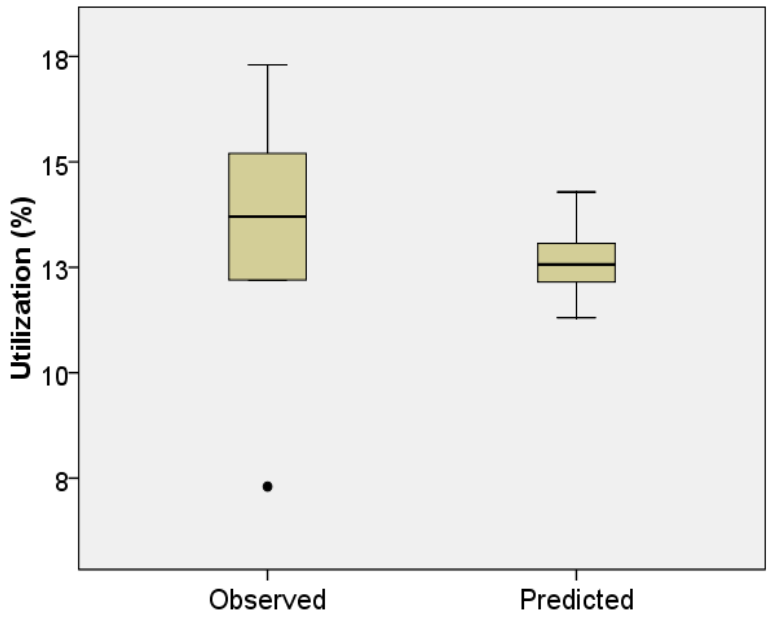


Figure 14: Observed and predicted utilization for best performers.

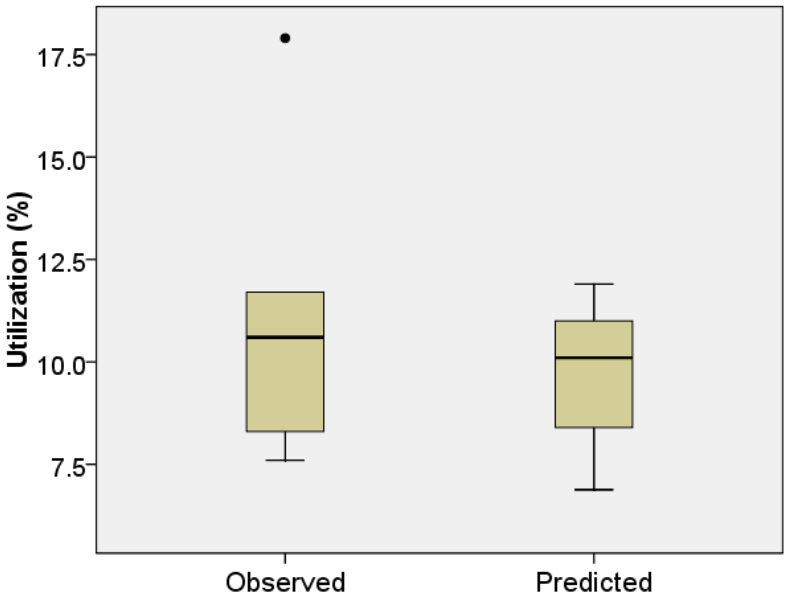


Figure 15: Observed and predicted utilization for worst performers.

4.2.3.3 Performance Scores

Replication of the observed performance scores from the human-in-the-loop study increases confidence in the model's ability to capture the effects of operator and automation performance (Nehme, 2009). The performance score represents the participants' ability to quickly find as many targets as possible and destroy as many hostile targets as possible. Figure 16 shows observed and predicted performance for all participants on a zero to two scale. The score represents operators' performance in quickly finding as many targets as possible and destroying hostile targets in a timely manner. The predicted performance score for each trial was computed according to Equation 3.5. The score was averaged over 500 trials. Average observed performance score for all participants was 1.13 (s.d. 0.4). The average predicted performance score was 1.04 (s.d. 0.14). Moreover, no statistical difference was established between the means

($t(24) = -0.09, p = 0.93$). The average observed performance score of best performers (Figure 17) was 1.69 (s.d. 0.07) and average predicted score was 1.70 (s.d. 0.03). The means were not statistically different ($t(4) = -0.40,\ p = 0.71$).

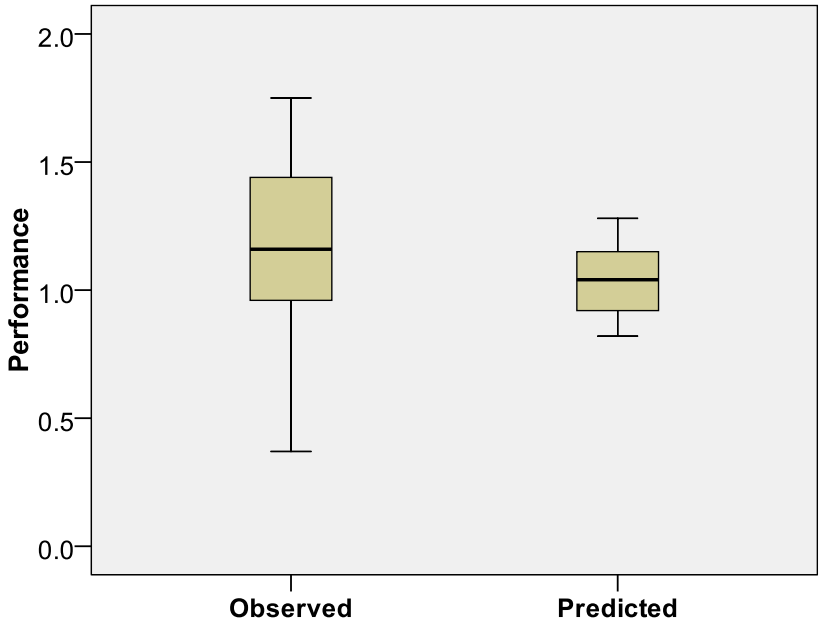


Figure 16: Observed and predicted performance for all.

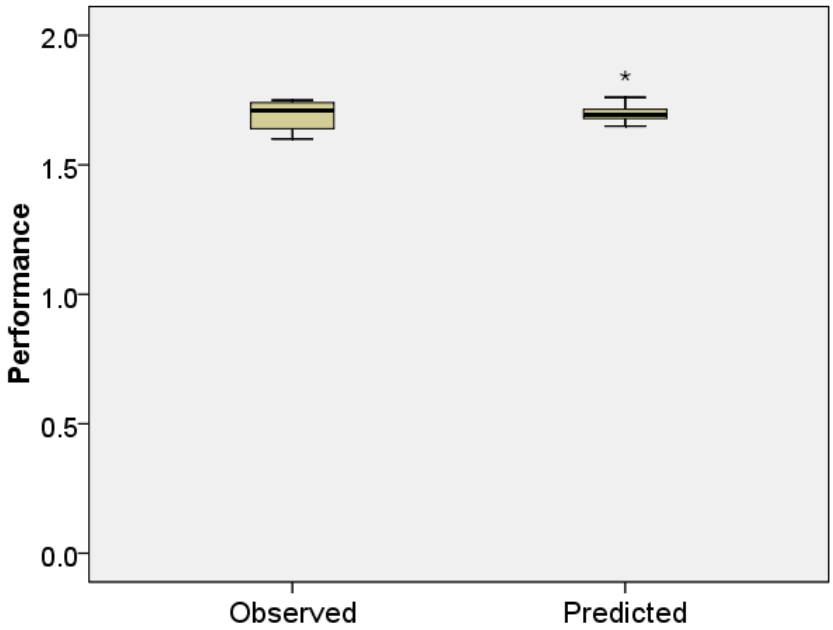


Figure 17: Observed and predicted performance for best performers.

In the case of the worst performers (Figure 18), the observed mean score was 0.55 (s.d. 0.11). The model predicted the average performance score to be 0.51 (s.d. 0.04) and there was no statistical difference compared to the observed value ($t(4) = -0.05, p = 0.96$). Overall, the model predicted performance scores were all within one standard deviation of the observed performance scores.

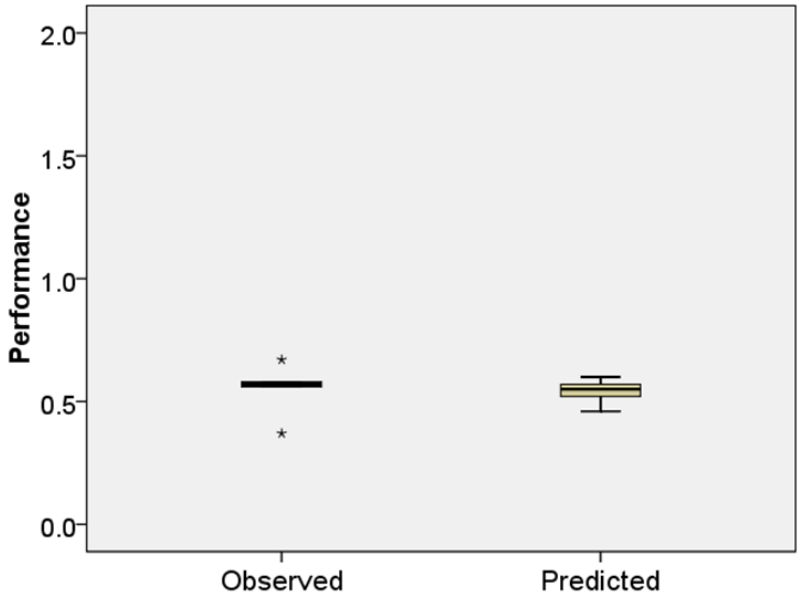


Figure 18: Observed and predicted performance for worst performers.

In all of the replications (i.e., number of events, utilization, mission performance score), the model yields adequately good results. Therefore, it is reasonable to conclude that the replication validity of the model can be established based on the historical data set of human-in-the-loop experimentation.

Besides the performance measures, the model can also be used to replicate observed attention switching patterns of operators. This is important in analyzing various attention switching patterns that can affect the system performance in different ways. This can also provide an

opportunity to evaluate whether an observed behavior is advantageous in improving system performance. In the next section, the model is utilized to simulate observed cyclical attention switching behavior.

4.3 Cyclical Attention Switching Strategy

The LTL-DES model can be used to simulate various attention switching behaviors of operators. Both the attention states and the switching pattern between the attention states can significantly affect the performance of the system. Undoubtedly, the performance of the system will be positively impacted if the operator supervising the system is in the directed attention state majority of the time. Similarly, if the operator is distracted the majority of the time, the system performance can be expected to be lowered because of higher probability to miss mission critical events. The data set of the low task load study (Hart, 2010) was used to investigate the effects of various attention states and attention switching strategies on system performance.

Figure 19 shows the observed attention states of the best performer in the low task load study. The best performer spent the majority of the time (90%) in the directed and divided attention states. It should not be surprising that in a low task load study, the operator who was distracted very little performed very well. Comparing the attention states of the second best performer (Figure 20) to the best performer (Figure 19), it can be observed that the second best performer spent about half of his time in the distracted attention state, but still managed to record a performance score only 1% behind the best performer. According to Figure 20, the percentage of directed attention state of the second best performer fluctuated in a cyclical manner. To further

analyze the effects of cyclical attention switching on operator performance, the LTL-DES model was utilized.

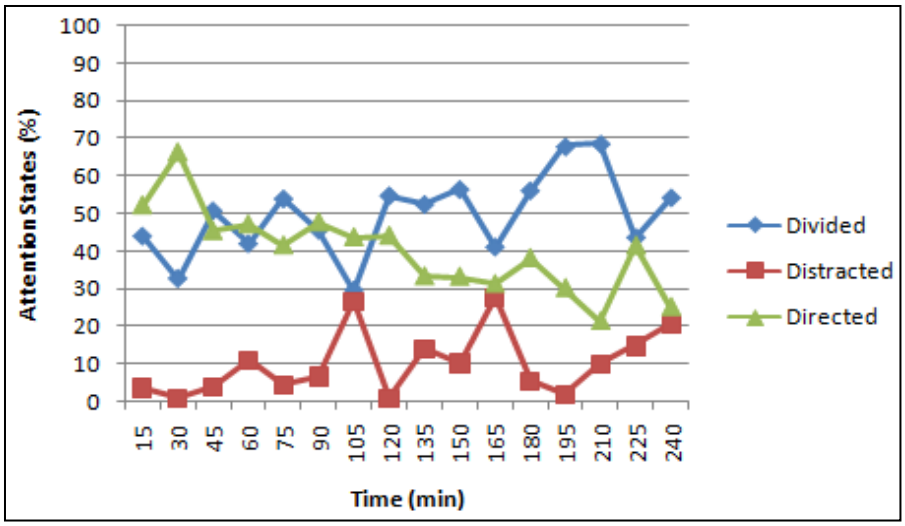


Figure 19: Observed attention states of the best performer.

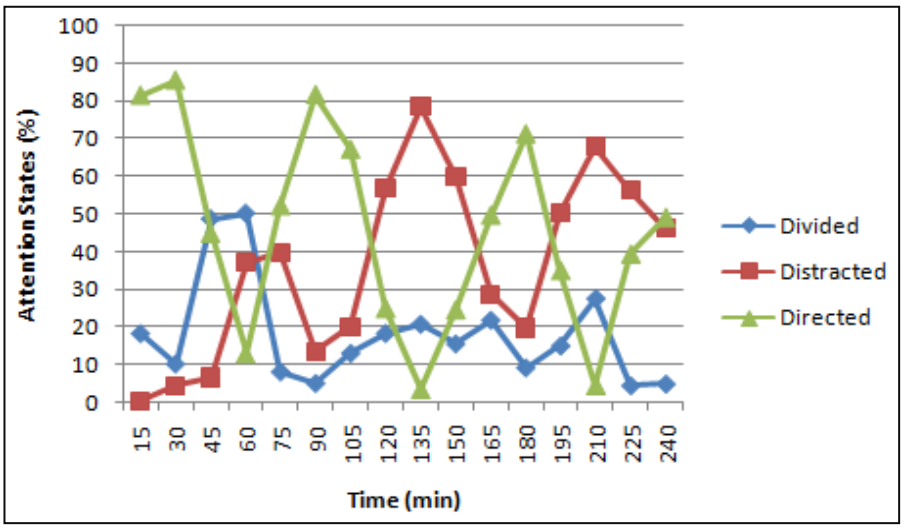


Figure 20: Observed attention states of the second best performer.

4.3.1 Effects of Cyclical Attention Switching on Operator Performance

As described in Chapter 2, sustaining directed attention in low task load supervisory settings can be challenging. Therefore, it is important to evaluate various attention switching strategies that

can help operators perform well. The cyclical attention switching strategy of the second best performer, described earlier, appeared to have a significant role in contributing to this participant's high performance score. To further assess the effects of cyclical attention switching strategies, the second best performer's directed attention state was approximated using a cosine function (Equation 4.1, Figure 21). Using this cosine function as an input to the LTL-DES, it is possible to evaluate the performance score of operators. According to the model predictions, the performance score of the worst performers should double if they switched their attention according to the cosine function (Figure 22).

$$f(t) = 40 + \left(40 - \frac{t}{24}\right) * \cos\left(\frac{t}{14 - \frac{t}{120}}\right),\ t \in [0,240] \tag{4.1}$$

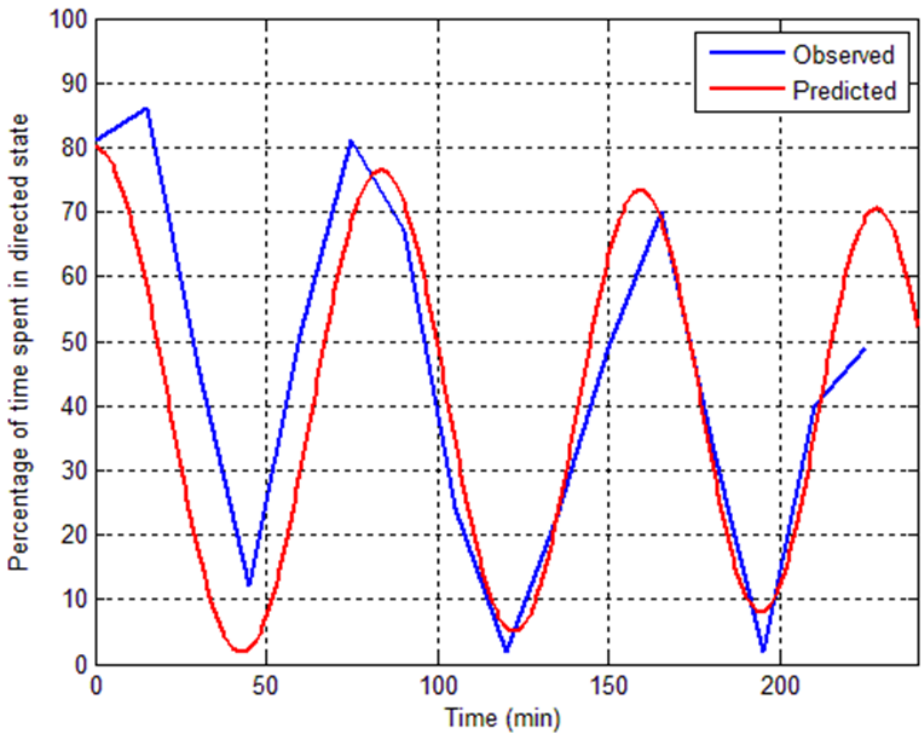


Figure 21: Observed and approximated directed attention state of the second best performer.

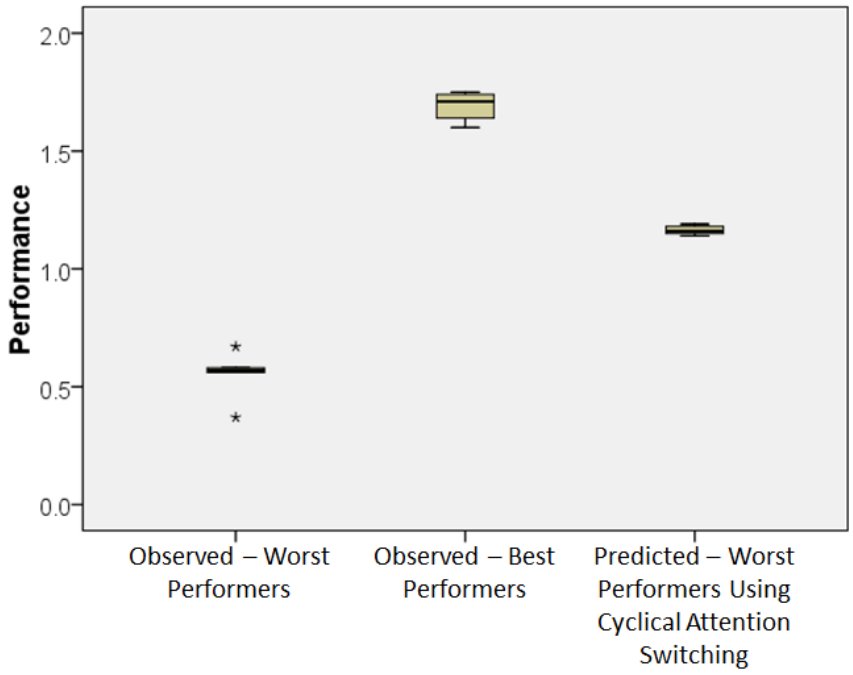


Figure 22: Predicted performance score of worst performers using cyclical attention switching strategy.

The results suggest that if it is possible to prompt operators of supervisory control systems to switch their attention in a cyclical manner, then dramatic performance increments could be achieved for those operators who generally do not perform well. It is important to note that prompting operators of supervisory control systems to switch in such a cyclical manner might be beneficial for the worst performers. However, cyclical attention switching strategy, most likely, will not be beneficial for the best performers, since majority of these operators monitor the system periodically, unlike the worst performers, who usually have hard time sustaining attention.

In order to evaluate the feasibility of prompting operators to pay attention to the system in a cyclical manner, a new long duration, low task load study was conducted. The study examined the possibility of improving system performance by utilizing the cyclical attention switching strategy. While the best performer spent most of her time directed and divided, it is unrealistic to

expect majority of operators to be distracted only 10% of the time in low task load domains. Hence, a possible design intervention based on the cyclical attention switching strategy was evaluated in the study. Additionally, the study was used for predictive validation of the LTL-DES, which is the next step in validating a simulation model. The details of the experiment and the results are presented in the next chapter.

5. Predictive Validation

A long duration, low task load study was conducted for two primary reasons. First, the results of the experiment were used for the predictive validation of the Low Task Load Discrete Event Simulation (LTL-DES) model. Second, the study was designed to evaluate the possibility and effectiveness of prompting operators of low task load supervisory systems to switch their attention in a cyclical manner. Before discussing the results and predictive validation of the model, the experimental setup is discussed.

5.1 Low Task Load, Long Duration Experiment

5.1.1 Apparatus

The simulation testbed used in this experiment, called the Onboard Planning System for Unmanned Vehicles Supporting Expeditionary Reconnaissance and Surveillance (OPS-USERS), allows a single human operator to supervise several highly automated UVs in a search, track, and destroy mission (Fisher, 2008). The interface was inspired by a futuristic UV control paradigm, in which a single operator is responsible for monitoring and controlling multiple UVs. The control structure is based on a high-level, goal-oriented scheme, rather than low-level, vehicle-based control. More specifically, operators are able to task the vehicles to search certain areas; however, they cannot specify altitude, heading, airspeed or other vehicle-level parameters. Instead, the operators specify locations on a map where they want the vehicles to travel to search, track, or destroy targets.

The main display of the OPS-USERS interface is the Map Display shown in Figure 23. The display shows a top-level view of the area in which vehicles are located. Operators are

responsible for supervising four UVs: (1) one vertical takeoff and landing unmanned aerial vehicle (UAV), (2) one fixed-wing UAV, (3) one unmanned surface vehicle (USV), and (4) one fixed-wing weaponized UAV (WUAV). Tasks are shown on the map as markers where vehicles need to go to explore. Targets can be (1) friendly, represented by blue color, (2) unknown, represented by yellow color, and (3) hostile, represented by red color. The symbols for the UVs and for the targets are based on DoD standards (US Department of Defense, 2008). The upper right corner of the Map Display shows a mini map, which is convenient to use when the Display Map is zoomed in and overall map view is not available.

The *Chat Message Box* is located in the lower right corner and depicts intelligence information. When a message comes in, participants hear a tone and the black outline of the box starts to blink. To stop the Chat Box from blinking, participants should click on the box.

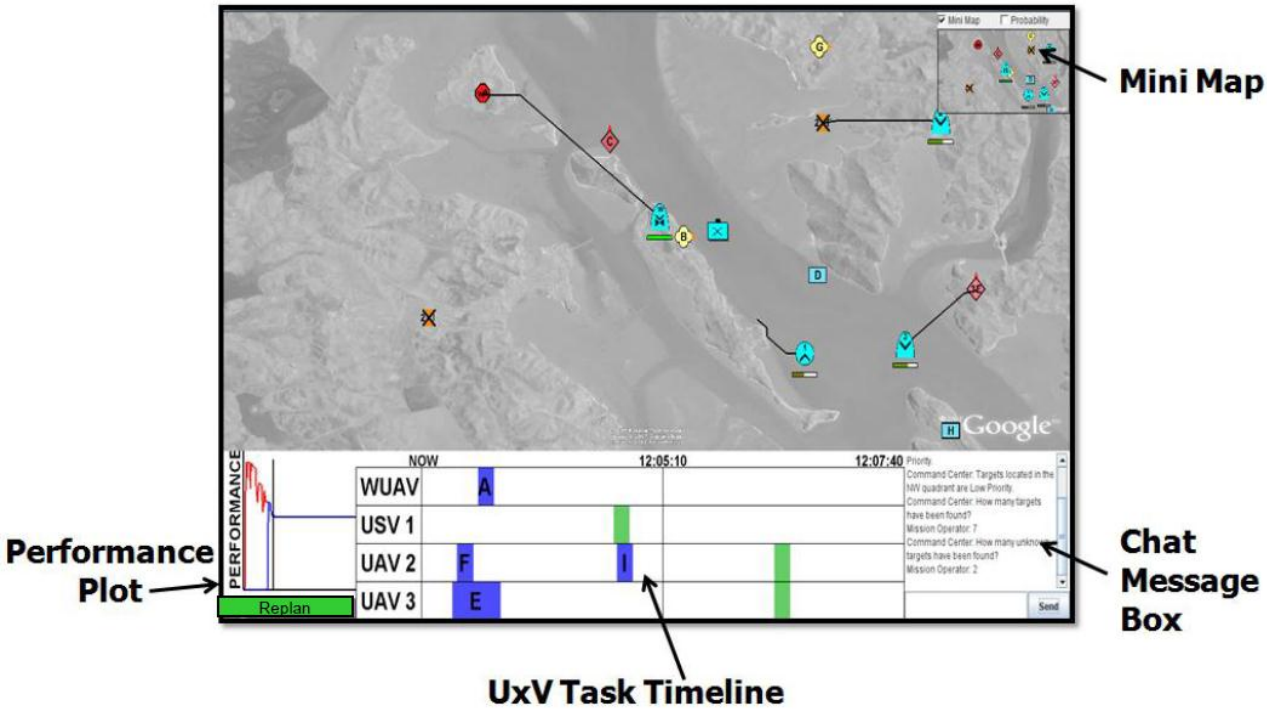


Figure 23: Map Display

The *Timeline* gives temporal information for each UV for the next thirty minutes, indicated in the format hours:minutes:seconds. Green bars in the Timeline indicate refueling, and blue bars indicate the UV is performing a task. The letter abbreviation of the task (whether Search Task or Target Track Task) appears in the blue bar. Each UV is limited to two task assignments at a time. White space indicates idle time or time traveling between tasks. The timeline shifts to the left as time passes.

The *Performance Plot* shows the auto-planner's computation of the current schedule's performance (red line) in comparison to the actual overall performance (blue line) of the current schedule. The plot moves to the left as time passes.

5.1.1.1 Operator Tasks

There are several tasks that operators need to complete over the course of the simulation: (1) replan, (2) create, edit, and delete search tasks, (3) identify targets, (4) approve weapons launch, and (5) respond to chat messages. These are discussed in detail below.

Replan

Since the vehicles operate in a dynamically changing environment, operators occasionally need to replan to update vehicles' path schedules. The Replan button in the lower left corner of the Map Display turns green when a new plan is available from the auto-planner, which is generated by a market-based, decentralized algorithm (Valenti, Bethke, Fiore, How, & Feron, 2006). Once the participant clicks the replan button, the Schedule Comparison Tool (SCT) appears (Figure 24) allowing operators to compare various path schedules for the vehicles.

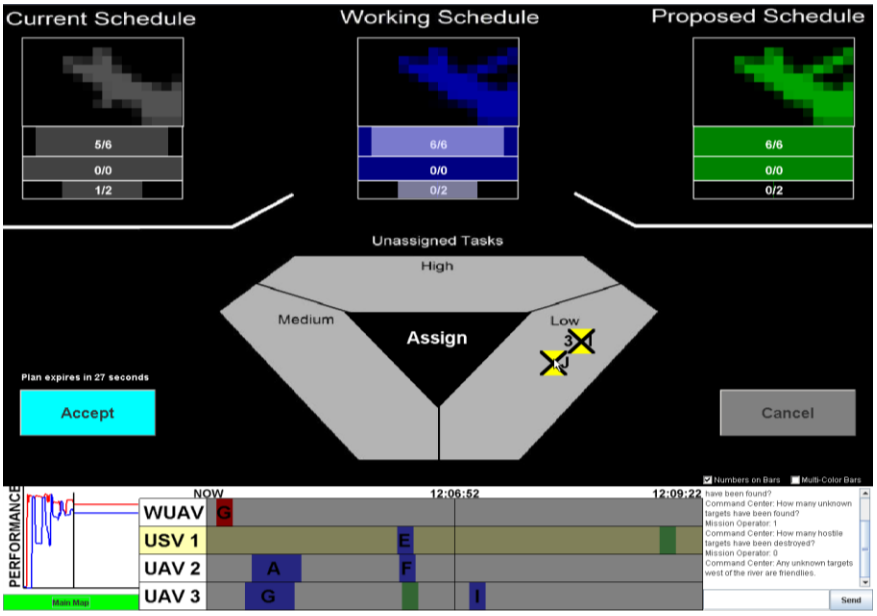


Figure 24: Schedule Comparison Tool (SCT).

Search Tasks

To command vehicles to search a specific area, operators need to create search tasks. Right clicking a location on the Map Display opens the Search Task Window shown in Figure 25. Right clicking an existing search task brings up the same window to edit the search task. The operator can designate the priority level of the task and the time frame in which it should be

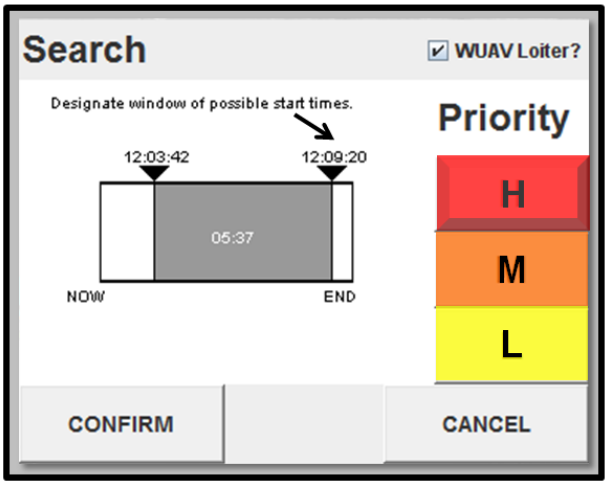


Figure 25: Search Task Window.

completed. If the checkmark “WUAV Loiter?” is selected, then the automated planner sends the WUAV to the search location.

Identify Targets

When UVs find targets, operator assistance is needed to identify these targets. In the OPS-USERS testbed, this process is simplified. When a target is found, a target identification window opens. Instead of actual imagery, symbols corresponding to either friendly, unknown, or hostile targets are displayed in the window. The operator pans through the target identification window until the target is in view, identifies the target, and assigns a priority level based on the mission parameters (Figure 26). The operator can later click on the same target to redesignate priority level or change the target status.

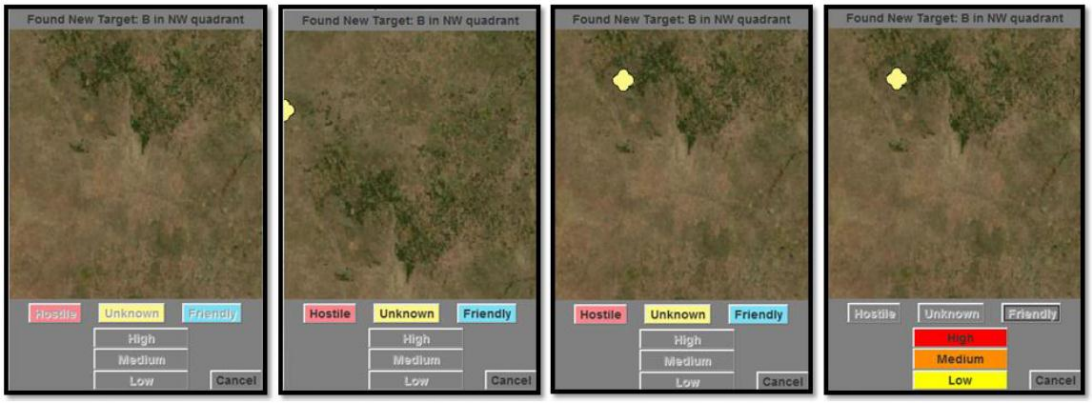


Figure 26: Target identification sequence.

Weapons Launch Approval

Before hostile targets can be destroyed, the operator has to approve the destruction of the target. The Missile Launch Approval Window pops up automatically (Figure 27) when the WUAV sights the hostile target for destruction and a second UV has the hostile in its sight. To destroy

the target, the operator pans through the screen for a direct view of the target and clicks the red Approve Launch button.

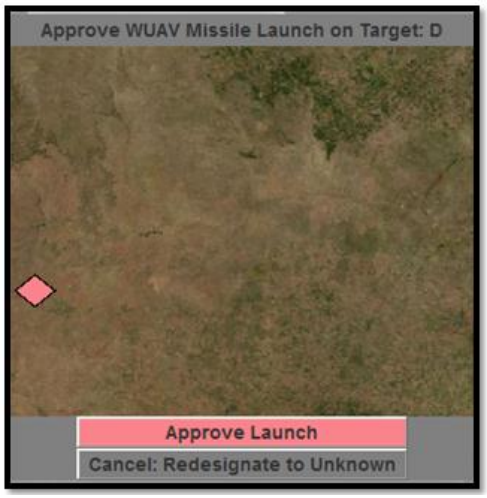


Figure 27: Missile Launch Approval Window.

Respond to Chat Messages

Occasionally, the operator receives messages from the Command Center. These messages can be informative (e.g., providing intelligence information), inquisitive (e.g., asking about the number of targets found), or commanding (e.g., asking to search a specific area). Operators are instructed to acknowledge the chat messages and implement an appropriate action as soon as possible.

5.1.1.2 Hardware

An operator workstation consisted of a Dell Inspiron desktop computer with a 17 inch monitor that was specifically allocated for the OPS-USERS interface. A second 17 inch monitor was available for the operators to use for non-simulation related purposes. The operators were videotaped using Microsoft™ HD web cameras for the duration of the experiment. One camera was allocated per operator and another camera recorded the overall view of the experimentation room. Lastly, all participants were required to wear wireless headphones, which allowed them to

move around the experimentation room and still be able to hear auditory alerts of the OPS-USERS interface. Participants were alerted every time a new chat message arrived, as well as when the system prompted to replan. The replan alert was implemented in the form of a computerized female voice pronouncing "Replan." The chat message alert consisted of two chimes that were approximately 100ms apart at 1400Hz and 1550Hz.

5.1.2 Participants

Nine participants, two females and seven males, operated the simulation three at a time to simulate a typical unmanned vehicle operating environment. Each participant had his own workstation running an independent version of OPS-USERS. Participants were compensated $400 for completing two four hour missions on different days. Also, they were informed that the person with the highest performance score would receive a gift card valued at $250.

Participants were recruited from the undergraduate and graduate student population of MIT. Ages ranged from 18 to 24, with mean of 20.7 years and standard deviation (s.d.) of 1.4 years. The participants had diverse video-gaming experience. More specifically, two participants specified that they played video games daily, while three participants specified that they rarely played video games. Also, the participants did not have any military experience (Appendix D).

5.1.3 Experimental Procedure

Before starting the experiment, participants read and signed the consent form. Next, participants completed a demographic survey, indicating their age, gender, occupation, military experience, gaming experience, sleep duration for the past two nights, and comfort level using computers. The NEO-Five Factor Index personality survey (Costa & McCrae, 1992) was administered, which rates participants' neuroticism, extraversion, openness to new experiences, agreeableness,

and conscientiousness. Lastly, the Boredom Proneness Survey (BPS) was administered (Farmer & Sundberg, 1986). The consent form and pre-experimental surveys are shown in Appendix E.

All participants completed a training session, consisting of a self-paced PowerPoint™ tutorial and a practice session using the OPS-USERS interface. The tutorial covered all functionalities of the interface, as well as the objectives of the operators. The practice session was an opportunity for the participants to further familiarize themselves with the interface and ask questions, if desired. The session lasted 15 minutes, after which participants took a five-minute break before starting the four-hour long test session.

During the test session, each participant was responsible for controlling four UVs. Overall, six targets were available to be found over the course of the experiment and half of the targets were hostile and needed to be destroyed. The targets were uncloaked every 40 minutes, starting at the tenth minute. Participants were allowed to interact with each other and use personal items, such as books, laptops and cell phones, though phone calls were not permitted. Additionally, snacks and a variety of non-alcoholic beverages were provided. All these items served as possible distractions from the OPS-USERS interface.

The experimenter remained in an adjacent room and monitored participants via webcams. Three times over the course of the study, the experimenter entered the test room to ensure that the simulation interface was working correctly and to check on the participants. If participants wanted to go to the restroom, the experimenter replaced them for the duration of the break, paused the simulation, did not interact with the simulation, and reported to participants any changes that had occurred.

After test session, participants completed a post-experiment survey, detailing their confidence level, busyness level, and the usefulness of auditory alerts on a five-point Likert scale, shown in Appendix E.

Participants completed two test sessions, starting between 10am and 1pm on separate days, and the post-experiment survey was administered after both test sessions. However, the training session was administered only prior to the first experimentation session. The two training sessions are discussed thoroughly in the next section.

5.1.4 Experimental Design

The study was conducted to evaluate the effects of the design intervention on operator performance in low task load supervisory domains. For this reason, each participant completed two test sessions: one with a design intervention to prompt cyclical attention switching and another test session without the design intervention. The order of the sessions was randomized and counterbalanced. Each participant completed no more than one session per day, starting between 10am and 2pm. The intervention was implemented in the form of auditory alerts that were pre-programmed in the interface. The alerts were designed to be distinct from all the existing aural alerts within the interface. The alerts consisted of four distinct chimes approximately 300ms long that resembled a doorbell sound. Between the first two and last two chimes there was a 400 ms pause. Between the second and the third chimes the duration of the pause was 1.2 seconds.

All participants were required to wear wireless headphones at all times to hear the alerts. The number of the alerts changed in a cyclical pattern (Figure 28) to promote the observed strategy, described in Chapter 4, by prompting participants to pay attention to the interface. Participants

were informed that the alerts indicated that the system needs operator attention. The time interval between the alerts for each 15 minute interval was designed by uniformly distributing the number of alerts shown in Figure 28 over 15 minute blocks.

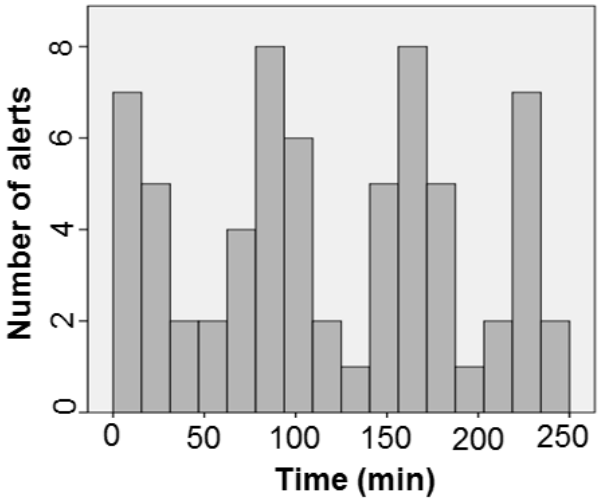


Figure 28: Number of auditory alerts shown in 15 minute blocks over the course of the experiment.

5.1.4.1 Variables

The independent variable in this experiment was the design intervention, resulting in two testing conditions.

The dependent variables were operators' workload, performance scores, participants' attention states, and subjective, self-rated metrics.

Workload (Utilization)

Utilization was used as a measure of objective workload. It is defined as percent busy time or the time operators spent interacting with various tasks divided by total available time. Although utilization does not account for the time that operators monitor the simulation, it is a useful metric that measures busyness level and has been used extensively to detect changes in workload (Cummings & Nehme, 2009; Cummings & Guerlain, 2007; Proctor & Zandt, 2008).

Performance Scores

The performance scores provide information on how well the objectives of the mission were accomplished (i.e., the speed and number of targets found and the speed and number of hostile targets destroyed). The overall performance score is comprised of the Target Finding Score (TFS) and the Hostile Destruction Score (HDS). The TFS accounts for the speed of finding targets and quantity of targets found. It is calculated as follows:

$$TFS = 1 - \frac{\sum_{i=1}^{N} \frac{t_i}{a_i} + (N_T - N)}{N_T} \tag{5.1}$$

where

t_i – time to find target i since it was available to be found

a_i – time target i was available to be found

N – number of targets found

N_T – total number of targets available

Equation 5.1 yields values between zero and one, where the higher the score, the better the participant performed.

In a similar manner, HDS is defined as:

$$HDS = 1 - \frac{\sum_{j=1}^{M} \frac{t_j}{a_j} + (N_H - M)}{N_H} \tag{5.2}$$

where

t_j – time to find jth hostile target since it was declared hostile

a_j – time target j was available to be found

M – number of hostile targets found

N_H – total number of hostile targets available

The HDS also ranges from zero to one, where the higher score indicates better performance. Summing the TFS and HDS yields a performance score with a range of zero to two.

Attention States

Operators' attention states were estimated by classifying their recorded activities. Three categories of attention states were identified: directed, divided, and distracted. In the directed attention state, the operator monitored the simulation interface or interacted with the interface. In the divided attention state, the operator monitored the interface while multitasking. An example can be eating while monitoring the interface. Lastly, in the distracted attention state, the operator did not pay attention to the interface at all (Appendix F).

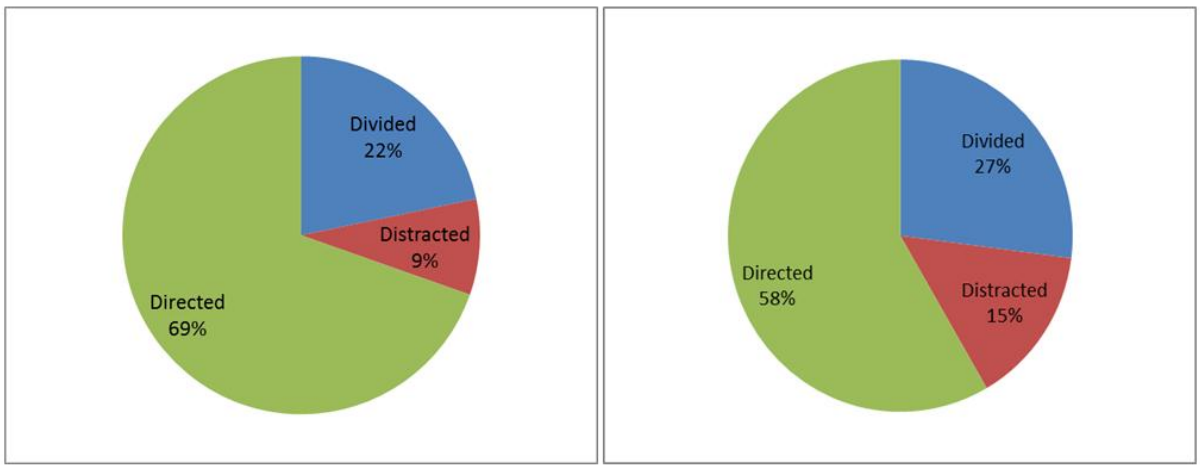
5.2 Experiment Results and Predictive Validation

Operators' utilization and performance scores were obtained by analyzing log files that captured operator interaction with the simulation interface. The recorded video files were used to approximate operators' attention states. Lastly, pre- and post-experiment survey data was analyzed to obtain subjective ratings. The following sections present the LTL-DES predictions and the results of the validation efforts. Before discussing model predictions, the analysis of attention states is presented, since the attention states serve as inputs to the model, as described in Chapter 3.

5.2.1 Attention States

In order to evaluate the attention states, two researchers watched the recorded videos and rated participants' attention states according to the same rule-based rubric (Appendix F). Overall, 18 four-hour-long videos were rated (coded): 2 four-hour videos per participant. The analysis revealed that in the first session, participants spent an average of 69% (s.d. 9%) of their time in a directed attention state, 22% (s.d. 4%) of their time in a divided attention state, and only 9% (s.d. 6%) of their time distracted. These results were very surprising, since in the previous low task load long duration experiment, described in Chapter 2, participants were directed 35% (s.d. 15%) and distracted about 44% (s.d. 20%) of their time. It should also be mentioned that the participants of this study were extremely similar in their interaction style, i.e., the majority of the participants stayed focused most of the time. The small standard deviation of the attention states also confirms this point. Figure 29a shows the average attention allocation of participants during their first session.

Similar results were obtained for participants in the second session (Figure 29b). More

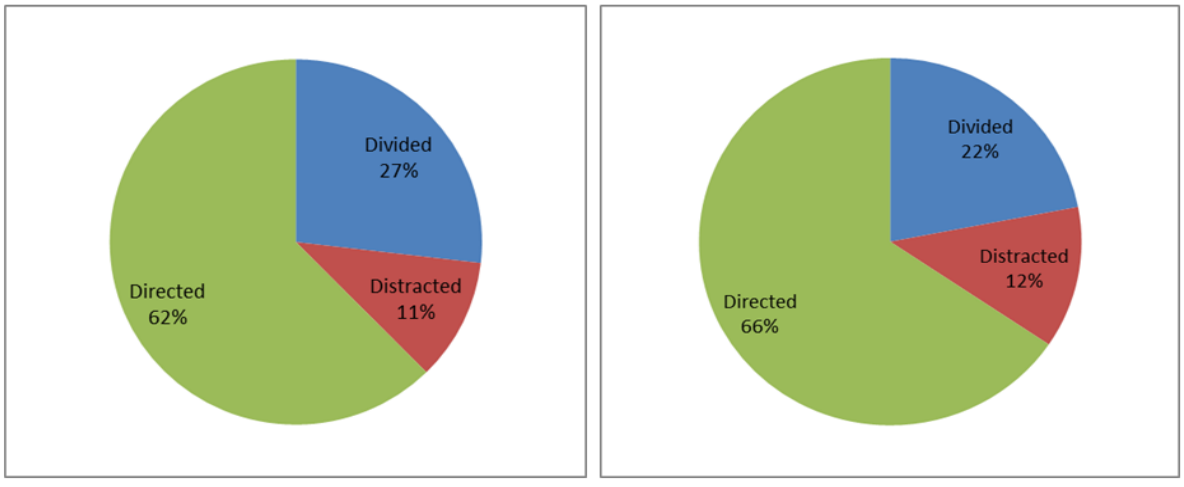


a. first session **b. second session**

Figure 29: Attention allocation of participants during first (a) and second (b) sessions.

specifically, participants spent on average 58% (s.d. 8%) in a directed attention state, 27% (s.d. 5%) in a divided attention state, and 15% (s.d. 6%) in a distracted attention state. Although the general trend was the same, i.e., participants were mostly directed and distracted very little, it is evident that participants became less directed during the second session ($t(8) = -4.33, p = 0.01$). As several participants mentioned in a post-experiment interview, after the first session they became more familiar with the interface and did not have to spend as much time monitoring the system to feel satisfied that they were achieving the objectives of the mission.

The allocation of attention states with and without the design intervention were also analyzed. Figure 30 illustrates the average attention allocation for both scenarios. In the scenario without the design intervention, participants spent an average of 62% (s.d. 7%) in a directed attention state, 27% (s.d. 4%) in a divided state, and 11% (s.d. 4%) in a distracted attention state. In contrast, with the design intervention participants spent on average 66% (s.d. 9%) of their time in a directed attention state, 22% (s.d. 6%) in a divided attention state, and 12% (s.d. 6%) of their time in a distracted attention state. It should be mentioned that the participant of the previous study who exhibited the innate switching strategy was only 35% directed on average. Also, the

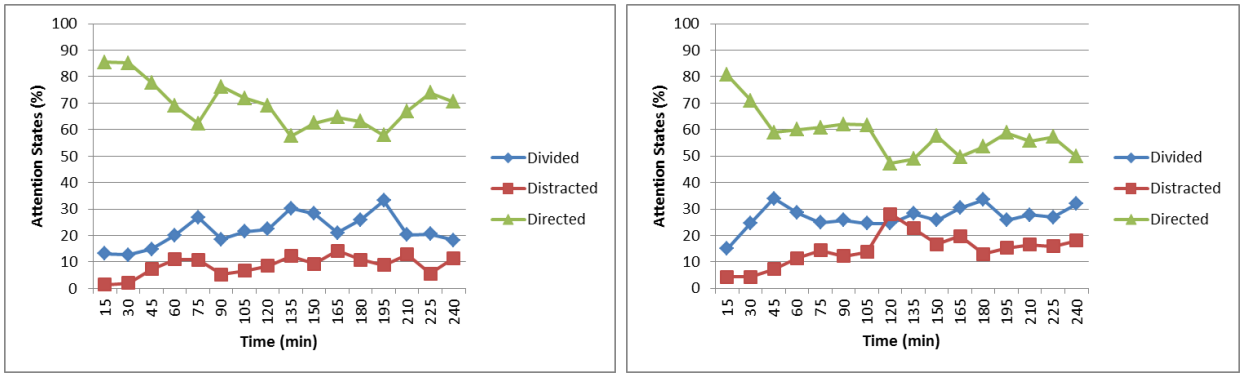


a. without intervention **b. with intervention**

Figure 30: Attention allocation of participants without (a) and with (b) intervention.

nearly equal proportions of attention states across the two scenarios (with and without the design intervention) indicate that the design intervention did not significantly affect the overall allocation of participants' attention resources. A paired t-test confirmed that no statistical difference exists ($t(8) = 0.71, p = 0.49$).

To evaluate whether the design intervention affected attention allocation over time, participants' attention states were analyzed in 15 minute increments. Figure 31 illustrates average attention states over the course of the four-hour-long study for the first and second sessions for all participants. Across the two sessions, participants were divided between 15 and 30 percent over the course of four hours. Also, during the first session, participants' directed attention declined from 85% to approximately 65% (on average), while during the second session the percentage of directed attention declined from 80% to about 55%. Hence, in the second session participants started less directed than in the first session and over time they became even less directed. This resulted in the percentage of distracted and divided attention states slightly increasing from the beginning of the study.



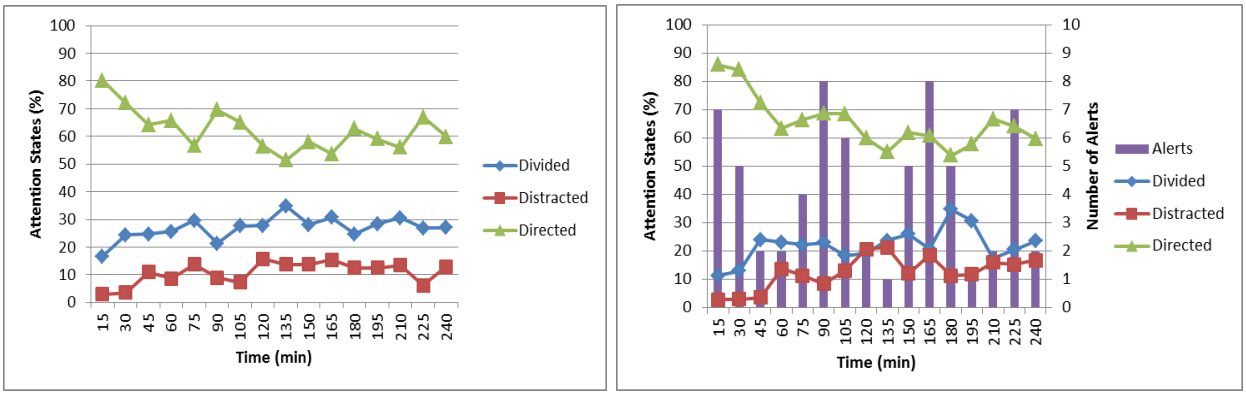
a. first session **b. second session**

Figure 31: Attention allocation over time during first (a) and second (b) sessions.

Figure 32 shows the percentage of attention states across the two scenarios, i.e., without (a) and with (b) the design intervention. In both cases, the percentage of directed attention declines about 25% in the first hour and then stays almost constant. The percentages of directed and distracted attention states, although fluctuate, do not increase or decrease significantly. Also, there are no statistical differences between the percentages of divided ($t(8) = -.178, p = 0.11$) and distracted ($t(8) = -0.67, p = 0.52$) attention states across the two scenarios.

Figure 5.29b shows several slight increases in the percentage of directed attention state, which appear to coincide with the maximum frequency of auditory alerts. However, the change in the percentage of the directed attention state is so small, that no definitive answer can be given on the cause of the fluctuations.

When comparing the percentage of the directed attention state and the number of alerts across all participants, there was a significant correlation only for one of the participants (Spearman's $\rho = 0.55, p = 0.03$). This participant's attention allocation over time is shown in Figure 33. It is interesting to note that this participant was the least directed among all the participants in the scenario with the design intervention. More specifically, he was directed on average about 40%



a. without intervention **b. with intervention**

Figure 32: Attention allocation over time during first (a) and second (b) scenarios.

of the time, divided 29%, and distracted 31% of the time. Therefore, the design intervention seemed to work for the most distracted participant, i.e. the participant switched his attention in a cyclical pattern. The design intervention appeared to work for this participant because he was not directed as much as the rest of the participants and was able to utilize the alerts to pay attention to the system. Furthermore, this participant's attention allocation was the most comparable to the attention allocation of the participant after who the design intervention was modeled (directed 37%, distracted 45%, divided 18%).

Appendix G summarizes the average percentage of attention states across the two sessions and two scenarios. Also, a table showing the average percentage of directed attention state in 15 minute blocks is presented in Appendix G, along with pairwise t-tests comparing percentages of directed attention states. No statistically significant differences were established across the scenarios.

The following two sections present model predictions of utilization and performance scores.

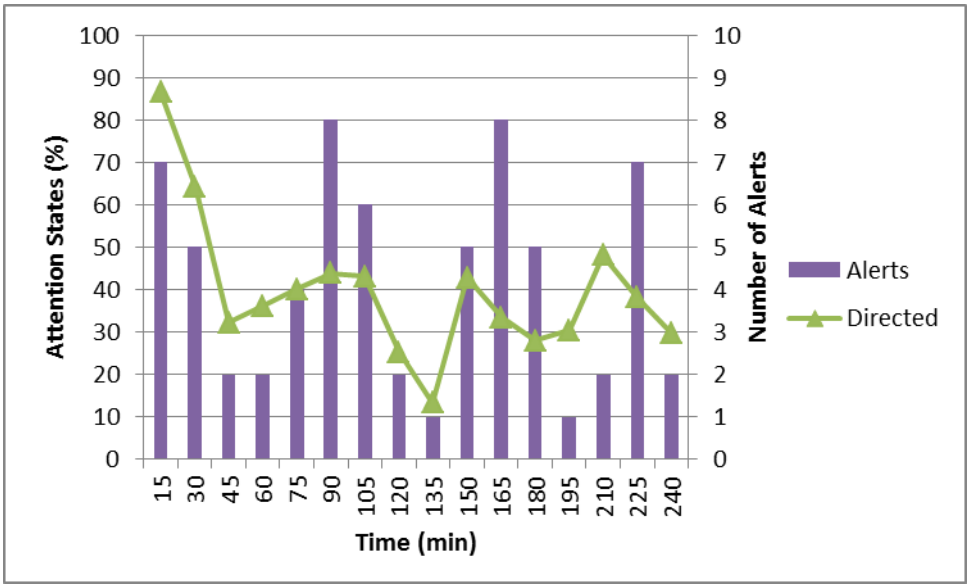


Figure 33: The least directed participant's directed attention state over time.

Since attention allocations across the scenarios were almost identical, only the attention allocation of the scenario with the design intervention was analyzed to serve as input to the model.

5.2.2 Utilization

The required average utilization of the study was about 2.1%, which was calculated based on the tasks that the system prompted operators to complete. However, total utilization was significantly greater than the required utilization. In fact, the average utilization in the scenario that did not include the design intervention was 14.5% (s.d. 4.8%) compared to 14.4% (s.d. 4.4%) average utilization with the design intervention.

To evaluate whether there were any differences between the first and second sessions (i.e., whether the order of the test sessions affected utilization), Table 1 shows the average utilization for the first and second sessions, as well as the average utilization with and without the design intervention. The results indicate that the design intervention did not affect the workload of the operators. It appears that during the second session, operators interacted less with the simulation interface. However, a within subject t-test showed that there is no statistical difference between the utilization in the first and second sessions ($t(8) = 1.72, p = 0.13$) and between the utilization of the scenarios with and without the design intervention ($t(8) = 0.06, p = 0.95$). Furthermore, there was no correlation between utilization in the first scenario and directed attention (Pearson's $\rho = -0.18, p = 0.64$) and utilization of the second scenario and directed attention (Pearson's $\rho = 0.34, p = 0.37$). Although one would expect participants who spent more time in the directed attention state to interact more with the interface, the results showed this was not true. The reason for this was that most of the participants who spent considerable

amount of time in the directed state did not necessarily interact with the simulation interface; rather, they monitored the interface.

Table 1: Utilization

	1st session	2nd session	With intervention	Without intervention
Average utilization (%)	15.0	13.9	14.4	14.5
Standard deviation (%)	4.2	4.8	4.4	4.8

The LTL-DES was designed to predict operator utilization. Participants' utilization was predicted to be 15.9% (s.d. 1.86%), based on 500 iterations of the model (Figure 34). Hence, the

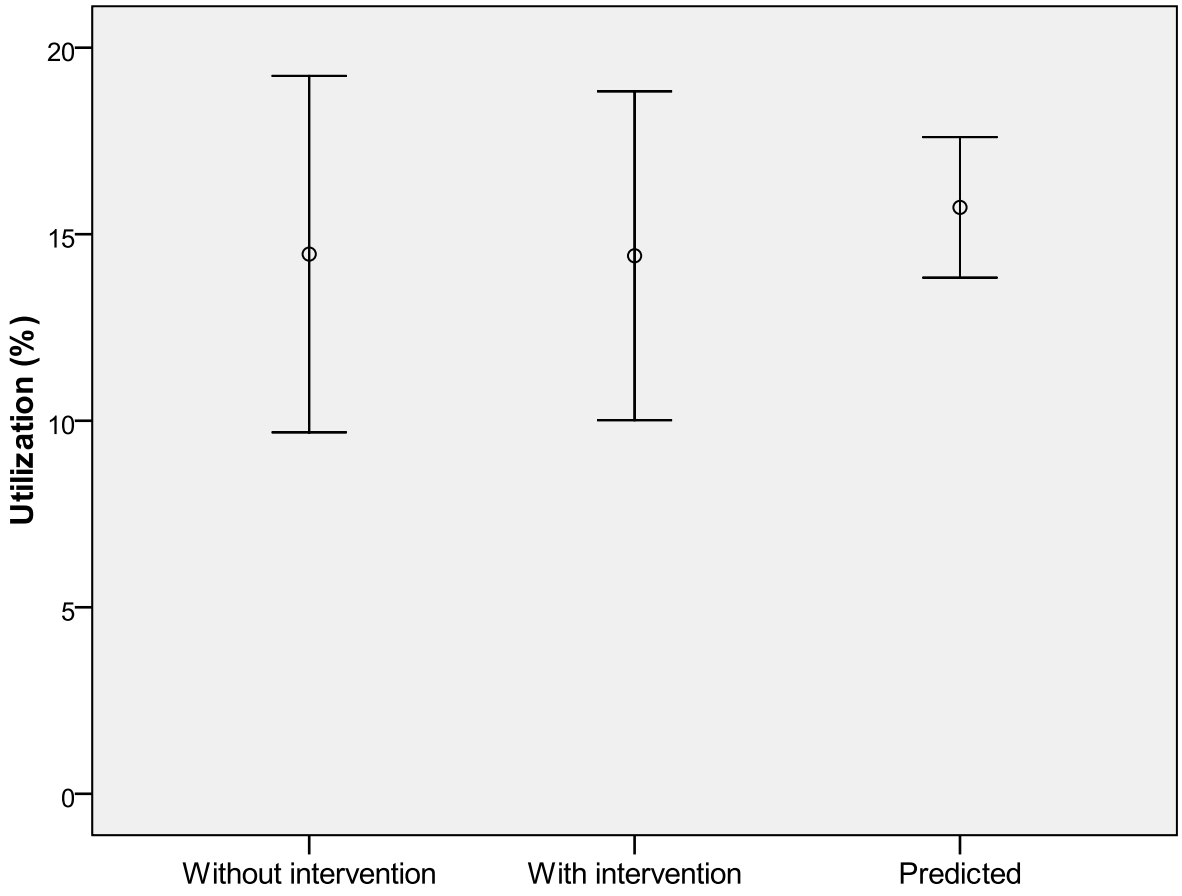


Figure 34: Mean and standard deviation of observed and predicted utilization.

observed utilization for both scenarios falls within one standard deviation of the predicted utilization. A t-test comparing the model prediction to the results of the scenario without the design intervention yielded no difference ($t(8) = -0.9, p = 0.29$). When comparing the prediction of the model to the results of the scenario with the design intervention, the results of the t-test also revealed no difference ($t(8) = -1.0, p = 0.34$). These results suggest the model is able to predict utilization in this low task load environment.

5.2.3 Performance Score

The performance scores were calculated based on Equations 5.1 and 5.2. The averages and standard deviations across sessions and intervention scenarios are shown in Table 2.

According to a within-subject t-test, there is no statistical difference across the sessions ($t(8) = -1.33, p = 0.22$) and the two scenarios ($t(8) = 1.87, p = 0.1$) using a 0.05 significance level. However, there seems to be a trend that indicates improvement in the performance score from the first session to the second session. This might be due to the learning effect, suggesting that as operators become more familiar with the interface, they understand it better and can make more effective decisions.

Table 2: Performance Scores

	1st session	2nd session	With intervention	Without intervention
Mean	1.56	1.72	1.53	1.74
Standard deviation	0.28	0.16	0.28	0.13

In contrast, the design intervention seemed to negatively impact participants' performance score, although there was no statistical difference. Most likely this was due to the fact that all of the

operators were directed most of the time and the auditory alerts were not necessary in prompting operators to pay attention to the interface.

To further validate the predictive ability of the LTL-DES model, the performance score was predicted. Figure 35 shows the predicted and observed performance scores. The predicted performance score was 1.78 (s.d. 0.08). As Table 2 shows, the average performance score of the scenario without the design intervention was within one standard deviation of the predicted performance score. A t-test revealed no significant difference between the means ($t(8) = -0.902, p = 0.39$). However, the average performance score of the scenario with the design intervention was statistically different from the predicted performance score ($t(8) = -2.66, p = 0.03$). The reason for the difference is the fact that during the second scenario, one of the participants, who was, on average, only 13% distracted, destroyed five targets, including two

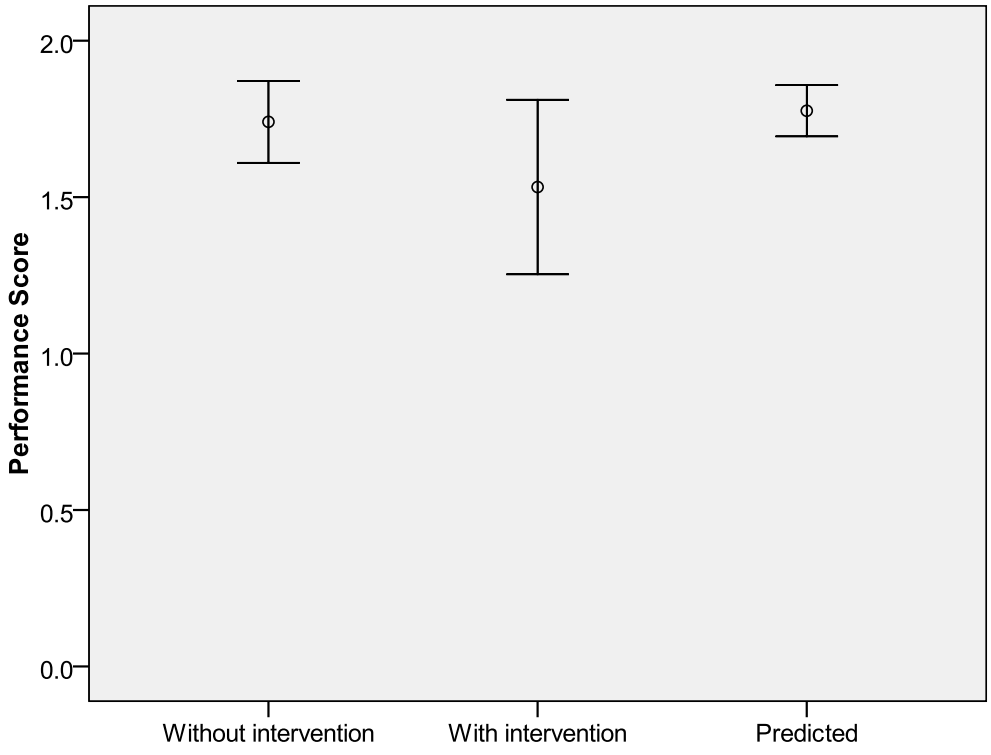


Figure 35: Observed and predicted performance scores.

friendly targets, significantly lowering his performance score. Also, another participant did not find one of the hostile targets and found other targets late, which lowered this participant's score. Nonetheless, this participant was, on average, only about 9% distracted.

5.2.4 Discussion of Predictive Validation

As described in the previous chapter, the LTL-DES model was developed to model operator performance in long duration, low task load supervisory domains. After the replication validation of the model, the next validation technique is the prediction validation, which was one of the goals for conducting the low task load, long duration study presented above.

The observed utilization of the participants was not statistically different from the predicted utilization; therefore, the model predictions of utilization were accurate. Compared to a previously-conducted similar study (Hart, 2010), utilization was about 3.5% higher. More specifically, the average utilization of the new experiment was 14.5%, while the average utilization of the previous experiment was about 11%. This means that participants of the new study had different interaction strategies compared to the participants of the previous study.

Also, the large range of utilization indicates the range of strategies that operators employed to supervise the system. More specifically, the highest utilization (22.9%) was about three times greater than the lowest observed utilization (8.3%). Interestingly, both the highest and the lowest utilization values were observed in the scenario without the design intervention.

The performance score prediction was not significantly different from the performance score results of the scenario without the design intervention. However, the prediction of the model was statistically different from the observed values of the scenario with the design intervention. The inability of the model to accurately predict the observed results is due to the fact that the model

does not take into account operator error. More specifically, the model does not account for destruction of friendly targets by operators who are mainly directed. Accounting for operator error can be accomplished by closely examining the error rate and assigning probabilities to making certain errors in judgment, which then can be implemented in the LTL-DES model.

5.2.5 Discussion of Design Intervention

The purpose of the design intervention in the experiment was to improve the performance of operators who had difficulties sustaining directed attention by prompting them to pay attention to the interface in a cyclical fashion. Unfortunately, in this study, there were not any participants who had difficulties sustaining attention. Moreover, all these participants performed very well, compared to the previously-conducted experiment. More specifically, in the previous study, the worst performers had a performance score of less than 0.7, while in the new study the lowest performance score was 1.0. Figure 36 illustrates the difference in performance scores. It should be noted that the performance scores across the two studies were normalized by taking into account various numbers of available targets. Equations 5.1 and 5.2 detail the normalization process.

In the new study, all of the participants were surprisingly directed, i.e., the percentage of their directed attention state was significantly higher compared to the previous study ($t(42) = -6.75, p = 0.00$). This can be considered the main reason the design intervention was not as effective in creating a cyclical attention state as hypothesized. In other words, prompting an operator who is already paying attention to the interface to pay attention in a cyclical fashion is not effective. However, the study showed that the participant who was directed the least (40%) responded to the design intervention in the first three hours of the experiment and exhibited a

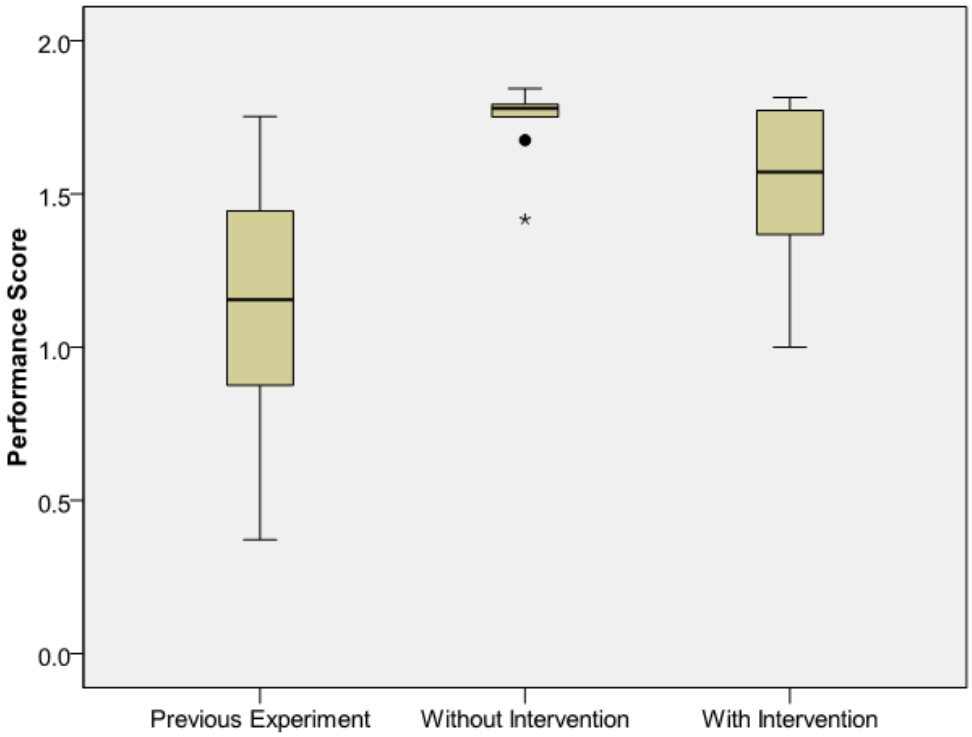


Figure 36: Performance scores of previously-conducted and new experiment.

cyclical attention switching pattern. This implies that more research needs to be conducted to assess the effectiveness of the design intervention on participants who are mainly distracted. To better understand the reasons the participants of the new study performed so well, demographic data of the participants, as well as pre- and post-experiment survey data were analyzed. This analysis can help extract personality traits to further refine the model by including other parameters that may affect performance. Subjective metrics are presented in the following section.

5.2.6 Subjective Metrics

Participants' self-rated metrics provide valuable subjective information on their perceived performance during the experiment on a five-point Likert scale, where zero is low and five is

high. Various subjective metrics were analyzed and are presented in the following sections. Appendix G shows descriptive statistics of these metrics.

5.2.6.1 Confidence Self-Rating

Participants rated their confidence level in the actions they took while interacting with the interface. Figure 37 illustrates all nine participants' self-rated confidence levels. Across the two sessions participants felt slightly less confident about the action they took in the scenario with

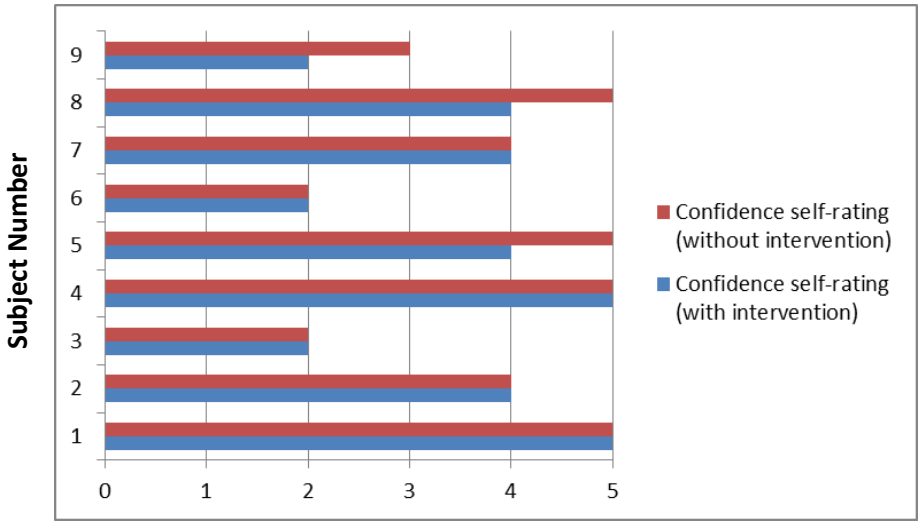


Figure 37: Confidence self-rating on a five-point scale.

the intervention. More specifically, the average confidence self-rating without the intervention was 3.9 (s.d. 1.3) and 3.6 (s.d. 1.2) with the intervention. However, there was no statistical difference between the scenarios with and without the design intervention using Wilcoxon Signed Rank test (Appendix G).

5.2.6.2 Performance Self-Rating

Across the two scenarios, all the participants specified the same self-rated performance level, except one participant (Figure 38). The average self-rated performance without the intervention

was 4.1 (s.d. 0.8) and 4.0 (s.d. 1.0) with the intervention. This means that the participants felt that they performed equally well with and without the intervention even though the two trials were conducted on different days.

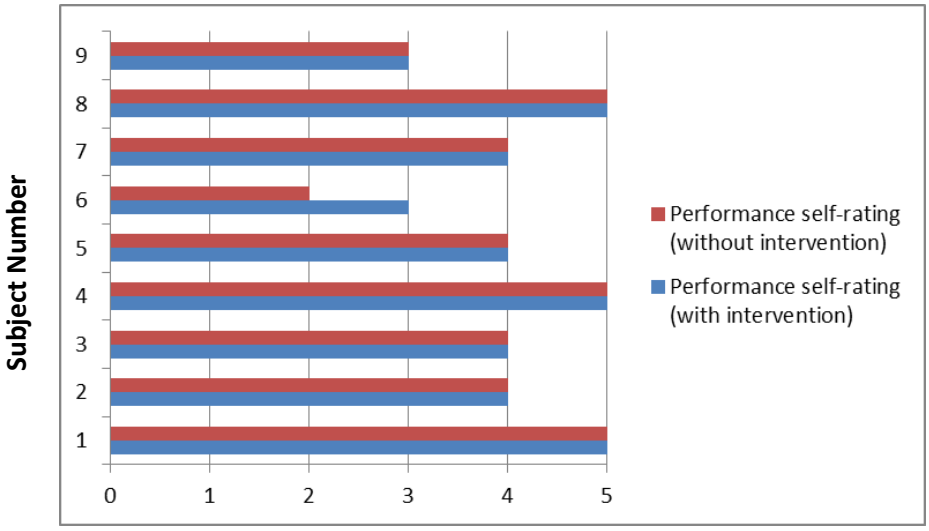


Figure 38: Performance self-rating on a five-point scale.

5.2.6.3 Busyness Self-Rating

When asked how busy participants were during the experiment, most of them answered (Figure 39) either *not busy* (2) or *busy* (3). Only twice participants mentioned that they were *idle* (1). None of the participants specified that they were *very busy* (4) or *extremely busy* (5). Also, two participants thought that the intervention caused them to work harder. These participants interacted with the interface less during the scenario without the design intervention compared to the scenario with the design intervention.

These results are similar to the self-rated busyness results of the previously-conducted experiment. More specifically, in the previous study, all the participants also rated their busyness between *idle* (1) and *busy* (3) (Hart, 2010).

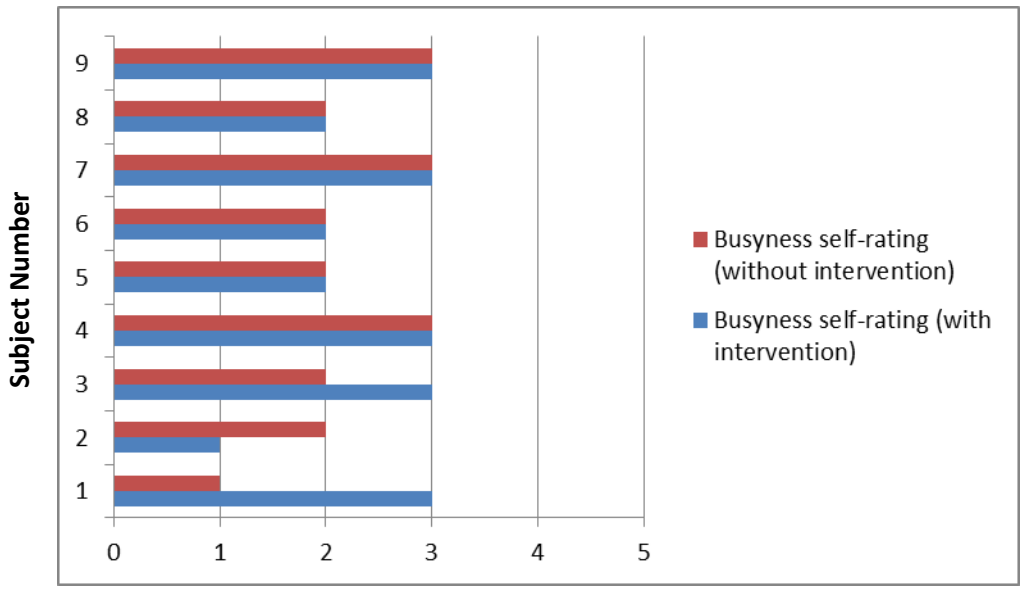


Figure 39: Busyness self-rating on a five-point scale.

5.2.6.4 Attention to alerts

Participants also rated how frequently they paid attention to the auditory alerts of the simulation interface (Figure 40). All of the participants answered either *frequently* (4) or *always* (5), which reflects the low distraction levels of operators.

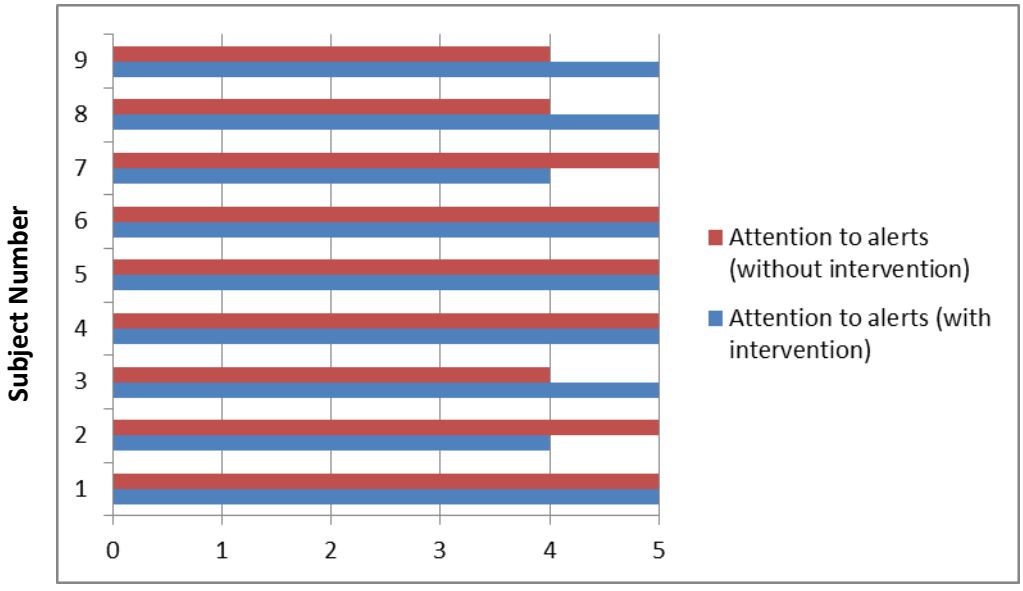


Figure 40: Attention to alerts self-rating on a five-point scale.

5.2.6.5 *Self-rated usefulness of alerts*

The usefulness of various alerts was also rated by the participants (Figure 41). A greater variability was observed compared to the frequency with which participants paid attention to the alerts. One of the participants specified that the alerts in the scenario with the intervention were *not useful* (1), while several other participants specified that the alerts were *extremely useful* (5). This variability means that participants utilized these alerts differently in their strategy of supervising UVs. It is important to note that the participant who specified that the alerts were *not useful* (1) was the most directed of all. This is logical, since this participant did not need the alerts to pay attention to the interface. It should also be mentioned that in the scenario without the design intervention only the replan and chat message alerts were implemented.

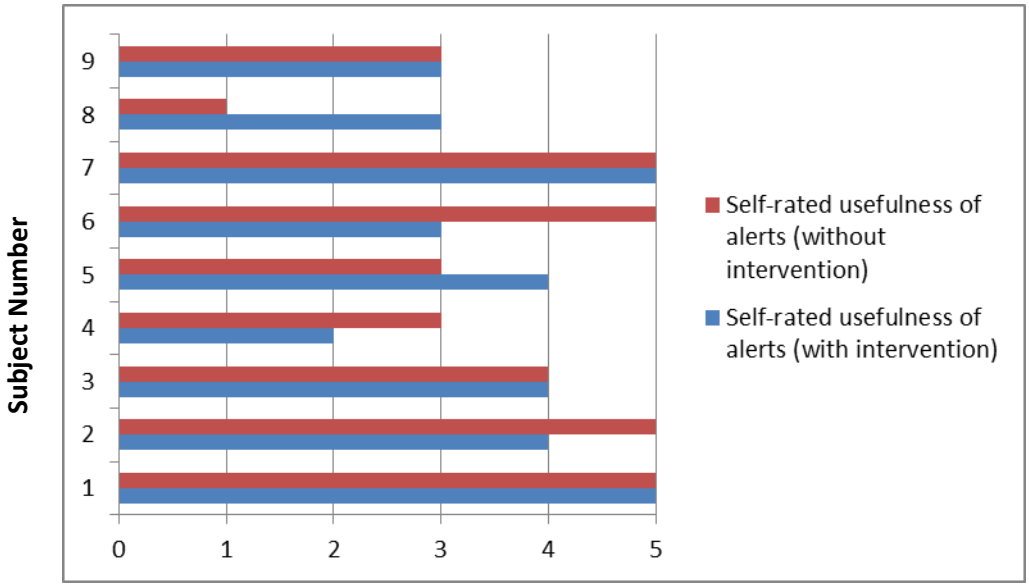


Figure 41: Self-rated usefulness of alerts on a five-point scale.

5.2.7 Personality Profiles

A NEO Five Factor Inventory psychological personality survey (Costa & McCrae, 1992) was administered to evaluate whether the personality of participants makes them more or less apt to

perform well in low task load supervisory conditions. The personality dimensions measured by the survey are neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. A description of the dimension characteristics is presented below:

1. Neuroticism – anxiety, hostility, depression, self-consciousness, impulsiveness, vulnerability to stress
2. Extraversion – warmth, gregariousness, assertiveness, excitement seeking, positive emotion
3. Openness to experience – fantasy, aesthetics, feelings, actions, ideas, values
4. Agreeableness – trust, straightforwardness, altruism, compliance, modesty, tender-mindedness
5. Conscientiousness – competence, order, dutifulness, achievement striving, self-discipline, deliberation

Figure 42 shows boxplots of the five personality dimensions discussed above. From the five dimensions, ratings for extraversion and openness to experience are significantly different from the theoretical mean (50), which implies that the participants who completed the study were significantly different from an "average" person. Appendix G provides personality dimension scores for all nine participants, along with t-test results comparing the observed ratings of the personality dimensions to the mean theoretical average of 50 (McRae & Costa, 2010).

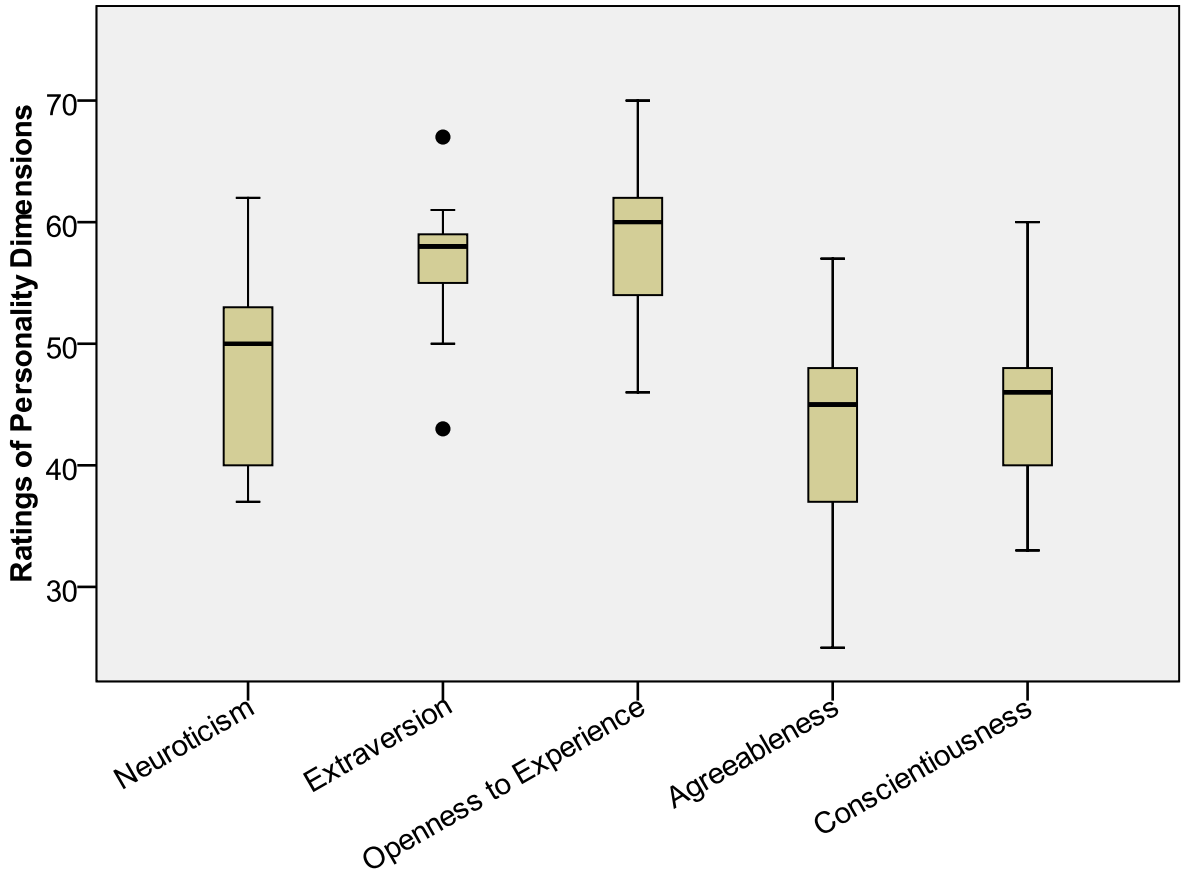


Figure 42: Boxplots representing five dimensions of the personality survey.

To evaluate whether performance scores were correlated with the personality dimensions, Spearman's correlation test was conducted. According to the results (Appendix G), no correlations were found between the personality dimensions and the percentage of participants' directed attention. Also, there were no strong correlations between the personality dimensions and performance scores of the two sessions. However, conscientiousness was marginally correlated ($Spearman's\ \rho = 0.65, p = 0.06$) with operators performance scores in the scenario without the design intervention. The conscientiousness score can possibly be a significant consideration factor when selecting future operators for low task load supervisory domains.

5.2.8 Boredom Proneness Scale (BPS)

The BPS measures people's propensity to become bored (Farmer & Sundberg, 1986). As described in Chapter 2, boredom can be a significant factor that impacts performance of operators in supervisory domains. The 28-item BPS was used to assess participants' boredom proneness levels. According to previously conducted studies (Winter, 2002), the sample mean of the US population is around 10.5. Also, participants who score below 5 are very low on the BPS, and those who score above 15 are very high on the scale. The results revealed that the majority of the participants had low boredom proneness levels (Appendix H). More specifically, the average BPS score was 7.8 (s.d. 4.0), minimum score was 4.0 and maximum score was 16.0 on a 28-point scale. Given the low BPS scores, it is not surprising that, on average, participants were only 12% distracted during the experiment.

To assess whether the BPS score could be used to predict operator performance, correlation coefficients between the BPS scores and the performance scores were calculated. The results indicate that no correlation exists (Appendix H). This is important, since it suggests that boredom proneness was not a major factor affecting participants' performance. In fact, the best and the worst performers specified the same level of boredom proneness. However, it might be that the homogeneity of participants (i.e., on average low BPS scores) contributed to the participants being mostly directed, which in turn resulted in high performance scores and no correlation between BPS and performance scores. In the future, a new study with a more diverse group of participants may yield significant results.

In the following section, subjective and objective data of the best and worst performers is analyzed in detail to better understand the differences between these participants that caused one

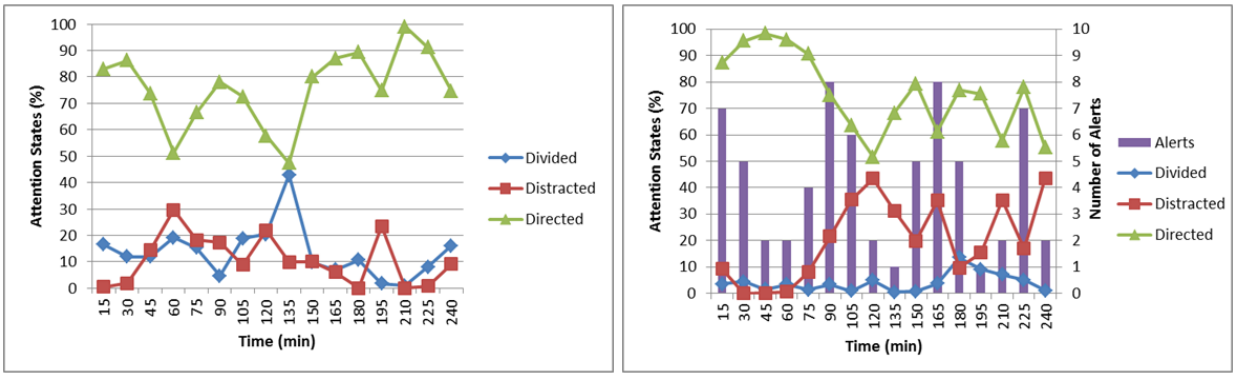
of them to perform very well and the other participant at about 50% of the theoretical maximum performance.

5.2.9 Best and Worst Performers Analysis

This section describes the behavior and performance of the best and worst performers. Their performance scores were calculated by summing the TFS and HDS of the first and second scenarios. The resultant score had a range of zero to four, the higher the better.

5.2.9.1 Best Performer

The best performer was a 21-year-old undergraduate student with no military experience, and who plays video games every day. His performance score was 3.64 and of all the participants who completed the study, he was the least prone to being bored (4 out of 28 on BPS). The ratings of the personality dimensions were not extraordinarily high or low. He also specified that the two days prior to both experimentation sessions, he slept at least seven hours. He self-reported a very high confidence level and performance level. This participant's self-reported busyness level was three (*busy*), self-reported attention to alerts was five (*always*) across the two sessions. Figure 43 shows this participant's attention states over time. Without the intervention, which was this



a. without intervention **b. with intervention**

Figure 43: Best performer's attention allocation over time.

participant's first session, on average he was 76% directed, 13% divided, and 11% distracted. In contrast, during the second scenario, he was 76% directed, 4% divided, and 20% distracted. It appears that the best performer relied on the design intervention to shift his attention from divided state to distracted state, while maintaining the same percentage of directed attention. It is also interesting to note that in the scenario without the design intervention the participant appears to have a cyclical attention switching strategy, although he was not prompted in that scenario to switch in a cyclical pattern.

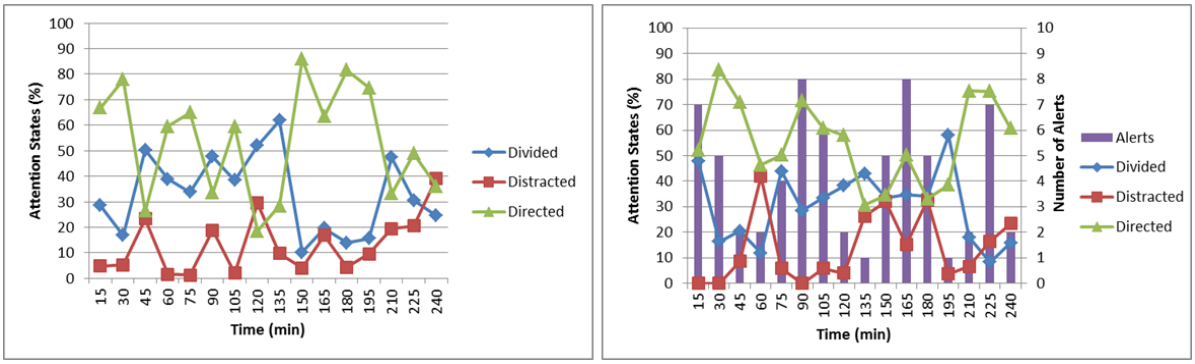
The observed utilization of the best performer was the second highest overall: 21.1% in the first scenario and 18.6% in the second scenario. It should be reiterated that the required utilization was only 2.1%; hence, the winner interacted with the interface almost 10 times more than required.

5.2.9.2 Worst Performer

The worst performer was a 20-year-old male with no military experience who also regularly played video games. His performance score was 2.78 because, in the second scenario, this participant destroyed two friendly targets. His propensity of being bored was also very low (4 out of 28 on BPS). The most interesting fact about this participant was his sleep schedule in the days prior to the experiment. He specified that the night before his second scenario he slept only 30 minutes and the day before slept 13 hours. Therefore, it should not be surprising that with only 30 minutes of sleep, this participant destroyed two friendly targets and rated his confidence and performance to be a four on a five-point scale. The personality survey ratings did not indicate any strong personality traits. This participant's self-reported busyness level was two (*not busy*) and self-reported attention to alerts was five (*always*) across the two sessions. Figure 44 shows

the attention states of the worst performer over time. Compared to the best performer, the worst performer was less directed, more divided, and distracted about the same. Also, the attention state pattern of the worst performer had greater variability compared to the top performer. Additionally, this participant spent considerable amount of time in the divided attention state by using his smartphone. Lastly, it is important to note that even though this participant performed the worst in the new low task load experiment, compared to the worst performer of the previous experiment, he performed much better. More specifically, the worst performer of the previous study had a performance score of 0.37, which is significantly lower than the lowest performance score (among the scenarios) of the worst performer (1.0) of the new study.

Unlike the best performer, this participant completed the scenario with the design intervention first and the scenario without the design intervention two days later. The observed utilization values of the worst performer across the scenarios were 8.3% and 11.4%, resulting in an average utilization of 9.85%, which was the lowest average utilization among the participants. Interestingly, there was no correlation between the average utilization of participants and their performance scores, despite the fact that the worst performer had the lowest average utilization



a. without intervention **b. with Intervention**

Figure 44: Worst performer's attention allocation over time.

and the best performer had the highest average utilization. The main reason for the lack of correlation was the large variance in performance scores and utilization among other participants. A table containing utilization values for all participants across the two scenarios is presented in Appendix H.

5.3 Summary

This chapter presented predictive validation of the LTL-DES model. A long duration, low task experiment was described, including the apparatus used in the experiment, as well as the required operator tasks. Next, participant information, experimental procedure, and the experimental design were discussed. The results of the experiment were analyzed and compared with the predictions of the LTL-DES model. This comparison revealed that predictions of the model with regards to the utilization of operators were accurate. However, the model accurately predicted performance score only in one of the design scenarios. More specifically, the model overestimated the performance score of operators who made errors by destroying friendly targets.

This low task load study also revealed that participants were significantly different from the general population in terms of their ability to sustain attention for prolonged periods. Furthermore, the design intervention implemented in the experiment to help operators of supervisory systems sustain directed attention could not be validated to have positive effects, mostly because the participants, in general, were highly directed. Lastly, it was established that, on average, participants had a low propensity of being bored. Over the course of the study, the

participants were distracted only about 12% of the time, which is remarkable given the very low task load nature of the experiment.

6. Conclusions

In highly automated, low task load supervisory control systems, sustaining focused attention for prolonged periods is challenging. Operators of these systems are often unable to find effective ways to combat consequences of boredom and vigilance decrements that negatively impact their performance. A tool for predicting operator performance is necessary to evaluate the effects of various design interventions and changes in system parameters oriented to improve operator performance. This research was motivated by the need to develop a simulation model to predict operator performance in low task load supervisory domains. Based on this model, a design intervention to help improve performance of operators who have difficulties sustaining attention was designed and evaluated by conducting a human-in-the-loop experiment.

The simulation model was based on discrete event simulation architecture and took into account different attention states of operators, as well as specific tasks that can be serviced when interacting with the supervisory system. The design intervention was implemented in the form of auditory alerts, the number of which varied over time in a cyclical manner. An experiment utilizing a video-game like supervisory interface was conducted to assess the impact of the design intervention on operator performance and to validate the human performance model.

6.1 Research Objectives and Findings

The primary objective of the research was to determine whether it was possible to develop a simulation model to predict operator performance in low task load supervisory domains. This objective was analyzed through the following methods:

- Prior work in modeling human performance and supervisory domains was reviewed to identify the direction that needed to be taken to develop the model (Chapter 2).
- A preliminary model was developed based on prior work and data from a previously-conducted low task load, long duration experiment. The model took into account the main variables that were thought to affect operator performance (Chapter 3).
- The model was verified and replicative validation was established via a historical data set (Chapter 4).
- Predictive validation was established by conducting a long-duration low task load experiment (Chapter 5).

The review of prior work in Chapter 2 motivated this research by revealing the difficulties experienced by operators of low task load supervisory control systems and the research gaps that exist in predicting operator performance in lieu of these difficulties. To fill in these gaps, a previously developed DES model of human-UV interaction served as a base model to develop a Low Task Load DES (LTL-DES) model. The major constructs of the LTL-DES were presented in Chapter 3, along with an explanation of how the model operated.

To establish replication validation of the model, the results from a previously conducted experiment in which a single human operator supervised a group of unmanned vehicles were used. By comparing model replications of utilization, number of completed tasks, and performance scores to the observed values, validity of the LTL-DES was established.

To further validate the model, a new long duration, low task load experiment was conducted in which operators supervised a group of UVs in a find, track, and destroy mission. Predictive validation was established by comparing the predicted and observed values of utilization,

performance scores, and the number of completed tasks. The predictions of the model revealed the model's weakness in capturing the full variability of human operators. More specifically, the variability in the observed utilization and performance scores was greater than the predicted values. Also, the way the model calculates the performance scores was based on the percentage of directed and divided attention states and average task wait times, causing the model to overestimate the scores of the participants who were distracted very little, but did not perform well. In the future, the model needs to take into account the probability of operator error (e.g., destruction of friendly targets) to alleviate the overestimation of performance.

The second objective of this research was to establish whether a design intervention could help improve participants' performance scores and whether predictions of the LTL-DES model were accurate.

The design intervention was designed to imitate the observed attention allocation of a participant who performed very well in the previously-conducted experiment. According to the LTL-DES model, if the design intervention was successful in prompting participants with difficulties in sustaining attention to switch attention in a cyclical fashion, the performance of these participants would improve dramatically. To assess the effectiveness of the design intervention, as well as validate the predictions of the model, the experiment described in Chapter 5 was conducted. The results of the experiment were somewhat unexpected, since none of the participants appeared to struggle to sustain attention. This was surprising because the previous study revealed the majority of participants had difficulties sustaining attention. For this reason, there was no conclusive evidence whether the design intervention could help improve struggling operators' performance or not. For the same reason, predictions of the model with regards to performance improvement could not be validated.

6.2 Recommendations and Future Work

Although the LTL-DES model shows promising results in predicting operator performance in low task load supervisory domains, future investigation is necessary to further evaluate the predictive ability of the model. Also, the effects of the design intervention should be further examined to fully understand its impact on operator behavior and performance.

Recommendations for future work are presented based on the research described in this thesis:

- As described in Chapter 5, participants of the low task load, long duration study were highly focused over the course of the experiment. This made the design intervention largely unnecessary, since the participants did not need any help sustaining directed attention. In the future, a new low task load, long duration study needs to be conducted with a new set of participants who have difficulties sustaining focused attention to fully evaluate the design intervention. Furthermore, a better selection process for participants needs to be developed to reduce the number of participants who do not have difficulties sustaining attention over prolonged periods of time.
- The performance score predictions of the model are based on the percentage of time operators spent in directed and divided attention states and on the average task wait time in the queue. This relationship yielded good results for the previously conducted low task load study. However, as the new low task load study showed, when extremely directed participants make errors, the relationship does not hold true. More analysis is needed to determine the appropriate metrics to use in predicting operator performance. One potential way to address this issue is to include a Bernoulli probability in the LTL-DES that models the probability of operator error.

- The LTL-DES model cannot provide real-time operator performance predictions because it relies on the attentions states of operators to predict performance. The option of predicting performance in real-time can be important in that it can help the operator of the supervisory system to adjust his behavior accordingly. In the future, adapting the model so that it can predict performance in real-time should be considered.
- The auditory alerts that were implemented in the experiment were set *a priori* and did not rely on operator performance or on parameters of the mission. Further analysis should be conducted to determine whether it is more appropriate to have auditory alerts based on operator interaction pattern, mission tasks, or other parameters that might help identify the "right" time for the intervention.

In conclusion, this research shows that it is possible to model human-system interaction in low task load supervisory domains by utilizing queuing-based DES theory. The attention allocation of operators and attention switching strategies appear to be critical factors in estimating performance of operators. Nevertheless, operator performance variability is very large and must be considered by designers of low task load supervisory systems.

This research also examined the effects of a design intervention to help improve operator performance, however, the results were not conclusive whether the intervention was beneficial or not. In the future, more research needs to be conducted to assess the effects of various design interventions on system performance.

Appendix A: Fatigue Models

The Two Process Model is at the core of many models that address fatigue and performance. The conceptual assumptions of the model include a linear interaction of homeostatic and circadian processes and an exponential sleep inertia component (Achermann, 2004; Borbely & Achermann, 1999).

The Interactive Neurobehavioral Model estimates neurobehavioral performance as determined by a linear combination of circadian, homeostatic, and sleep inertia components (Jewett & Kronauer, 1999). The predictions of this model include minimum core body temperature, alertness level, and performance. The model has been used by NASA and DoD researchers for shift/duty scheduling purposes.

The System for Aircrew Fatigue Evaluation (SAFE) Model was developed mainly for use in aviation processes (Belyavin & Spencer, 2004). The model can be described as a combination of a cubic trend in time since sleep and a sinusoidal component in time of day. It has been reported that UK Civil Aviation used the SAFE model to predict alertness levels during duty periods of air traffic controllers (Mallis, Mejdal, Nguyen, & Dinges, 2004).

The Circadian Alertness Simulator allows for the assessment of fatigue risk based on sleep-wake patterns (Moore-Ede et al., 2004). This model is based on the Two Process Model. The output includes plots of predicted alertness as a function of time and the impact of naps at different times of day. The model does not account for different types of work or other stressors that may influence fatigue or alertness.

The Dynamic Bayesian Network Real-Time Fatigue Model is based on a hierarchical Bayesian network and takes into account temperature, light, anxiety, workload, head tilt, gaze, and yawn frequency to predict accumulation of fatigue over time (Lan, Ji, & Looney, 2003).

Appendix B: Queuing Notation

This notation was developed more than a half century ago (Kendall, 1953) and is still used extensively in describing the diverse array of queuing systems. The notation is based on the *A/B/c/N/K/D* format in which the letters represent the following characteristics of the system:

- *A* represents the inter-arrival time distribution.
- *B* represents the service time distribution.
- *c* represents the number of parallel servers. In the LTL-DES, the server represents the operator who services tasks.
- *N* represents the system capacity.
- *K* represents the size of the calling population, i.e., the number of events that can be generated.
- *D* represents queuing policy

The most common distributions for *A* and *B* are *M* (Markov or exponential), *D* (deterministic), E_k (Erlang of order *k*), *H* (hyper-exponential), *PH* (phase-type), and *G* (general). Occasionally, the last three letters (*N, K, D*) will be dropped from the description of the queuing system if the system has unlimited capacity and calling population. In some cases, such as *M*/*M/1*, *M/M/∞,* or *G/M/1*, the queuing systems can be analyzed mathematically to yield a limited number of useful performance metrics. However, in more general cases (e.g., *G/G/1* or *G/G/∞*), mathematical analysis cannot yield useful results, thus simulation models are used to analyze these systems.

Although it is possible for a human supervisory system to be so simple that it can be analyzed mathematically, almost always these systems are so complex that only simulation methods can provide information on the system performance. The LTL-DES is considered to be a *G/G/1* queuing system, since probability distributions characterizing inter-arrival and service types are not restricted to the limited number of popular distributions mentioned above. Also, since the LTL-DES models situations in which a single operator is involved in operating the system, the number of servers is one. Furthermore, as described in section 3.3, each event type in the LTL-DES can have a different inter-arrival distribution and arrival processes can be dependent. All of these, make finding a mathematical solution for the *G/G/1* system extremely challenging, if not impossible.

Appendix C: Service and Arrival Time Distributions

Distributions are computed using the log files of the previously-conducted low task load experiment (Hart, 2010), as well as the new experiment described in Chapter 5.
EasyFit™ software package was used.

C.1 Service Time Distributions

Event		Previous Study		New Study	
		Distribution	**Parameters**	**Distribution**	**Parameters**
Replan	Hour 1	Lognormal	Mu=1.2064 Sigma=0.3549	Lognormal	Mu=1.1381 Sigma=0.479
	Hour 2	Lognormal	Mu=0.946 Sigma=0.3341	Lognormal	Mu=1.0283 Sigma=0.3724
	Hour 3	Weibull	Scale=2.5938 Shape=3.3935	Lognormal	Mu=1.0625 Sigma=0.2951
	Hour 4	Weibull	Scale=3.3653 Shape=4.1927	Lognormal	Mu=1.1732 Sigma=0.3162
Create/Edit Search Task	Hour 1	Lognormal	Mu=1.1825 Sigma=0.235	Gamma	Shape=9.7187 Scale=0.3166
	Hour 2	Lognormal	Mu=0.86791 Sigma=0.2024	Gamma	Shape=9.5622 Scale=0.3051
	Hour 3	Lognormal	Mu=0.8092 Sigma=0.1463	Lognormal	Mu=1.0719 Sigma=0.2561
	Hour 4	Lognormal	Mu=0.7271 Sigma=0.1514	Gamma	Shape=10.411 Scale=0.2558
Read/Respond to Chat Message	Hours 1-4	Normal	Mu=5 Sigma=1	Normal	Mu=5 Sigma=1
Target Identification	Hours 1-4	Normal	Mu=7.4265 Sigma=2.8619	Lognormal	Mu=1.6954 Sigma=0.3315

C.2 Arrival Time Distributions

Required Events

Event	Previous Study		New Study	
	Distribution	**Parameters**	**Distribution**	**Parameters**
Replan	Normal	Mu=1200 Sigma=5	Normal	Mu=1200 Sigma=5
Add Task	Uniform	a=800 b=1500	Uniform	a=900 b=2100
Chat Message	Uniform	a=900 b=1500	Uniform	a=600 b=1300
Target Identification	Uniform	a=600 b=900	Uniform	a=500 b=1200

Self-imposed Events

Event	Previous Study		New Study	
	Distribution	**Parameters**	**Distribution**	**Parameters**
Replan	Gamma	Shape=7.4276 Scale=6.618	Lognormal	Mu=3.8481 Sigma=0.4055
Add Task	Lognormal	Mu=4.2035 Sigma=0.4586	Lognormal	Mu=3.9771 Sigma=0.3506
Edit Task	Lognormal	Mu=7.0593 Sigma=0.341	Lognormal	Mu=6.2094 Sigma=0.7461

Appendix D: Participant Information

Subject	Age	Gender	Military Experience	Gaming Experience [1 (low) – 5 (high)]
1	23	M	N	1
2	20	M	N	3
3	21	F	N	2
4	21	M	N	5
5	20	M	N	5
6	20	F	N	1
7	18	M	N	4
8	21	M	N	4
9	22	M	N	1

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age	9	18	23	20.67	1.414
Gaming Experience	9	1.0	5.0	2.889	1.6915
Valid N (listwise)	9				

Appendix E: Pre- and Post-experiment Forms

E.1 Consent Form

CONSENT TO PARTICIPATE IN

NON-BIOMEDICAL RESEARCH

Assessing the Impact of Cyclical Attention Switching Strategies in Controlling Multiple Unmanned Vehicles

You are asked to participate in a research study conducted by Professor Mary Cummings Ph.D, from the Aeronautics and Astronautics Department at the Massachusetts Institute of Technology (M.I.T.). You were selected as a possible participant in this study because the expected population this research will influence is expected to contain men and women between the ages of 18 and 50 with an interest in using computers. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

- **PARTICIPATION AND WITHDRAWAL**

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise which warrant doing so.

- **PURPOSE OF THE STUDY**

The purpose of this research is to study how different attention switching strategies affect performance in long duration, low workload scenario in the context of piloting multiple, highly autonomous unmanned vehicles.

- **PROCEDURES**

If you volunteer to participate in this study, we would ask you to do the following things:

- Participate in training on the video game-like interface using a PowerPoint tutorial, complete a fifteen-minute practice session where you will control a team of simulated unmanned vehicles (UVs). The vehicles you will control will be assigned with the task of finding, identifying, and tracking targets in an area of interest, destroying hostile targets, and collaborating with the auto-planner to replan schedules.
- Participate in two four-hour long testing sessions where you will experience a long duration, low workload mission. You may complete only one mission per day. You will work alongside two other participants to simulate a populated control room, and you will each have your own workstations with your own vehicles and territory to control. You will be required to wear a wireless headset to receive aural alerts from the interface.

- You will be rewarded a score for the each session based on the number of targets you successfully find, how long they are successfully tracked thereafter, the percentage of the total area of interest is searched, and the time spent to find hostile targets and destroy them.
- All testing will take place at MIT in room 35-220.
- Total time: 8 hours and 45 minutes
-

- **POTENTIAL RISKS AND DISCOMFORTS**

There are no anticipated physical or psychological risks in this study.

- **POTENTIAL BENEFITS**

You will gain a deeper sense of appreciation on how much human-UV interaction future unmanned systems might require. Also, the results from this study will assist in the design of human-UV interfaces.

- **PAYMENT FOR PARTICIPATION**

You will be paid $150 to participate in the first session and $250 for the second session. You will be paid upon completion of your debrief. Should you elect to withdraw in the middle of the study, you will be compensated for the hours you spent in the study. An additional $250 Best Buy Gift Card will be awarded to the participant with the highest score.

- **CONFIDENTIALITY**

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law. You will be assigned a subject number which will be used on all related documents to include databases, summaries of results, etc.

- **IDENTIFICATION OF INVESTIGATORS**

If you have any questions or concerns about the research, please feel free to contact the Principal Investigator, Mary L. Cummings, at (617) 252-1512, e-mail, missyc@mit.edu, and her address is 77 Massachusetts Avenue, Room 33-311, Cambridge, MA, 02139. The investigator is Armen Mkrtchyan. He may be contacted at (617) 253-0993 or via email at armen@mit.edu.

- **EMERGENCY CARE AND COMPENSATION FOR INJURY**

If you feel you have suffered an injury, which may include emotional trauma, as a result of participating in this study, please contact the person in charge of the study as soon as possible.

In the event you suffer such an injury, M.I.T. may provide itself, or arrange for the provision of, emergency transport or medical treatment, including emergency treatment and follow-up care, as needed, or reimbursement for such medical services. M.I.T. does not provide any other form of compensation for injury. In any case, neither the offer to provide medical assistance, nor the actual provision of medical services shall be considered an admission of fault or acceptance of liability. Questions regarding this policy may be directed to MIT's Insurance Office, (617) 253-2823. Your insurance carrier may be billed for the cost of emergency transport or medical treatment, if such services are determined not to be directly related to your participation in this study.

- **RIGHTS OF RESEARCH SUBJECTS**

You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143B, 77 Massachusetts Ave, Cambridge, MA 02139, phone 1-617-253 6787.

SIGNATURE OF RESEARCH SUBJECT OR LEGAL REPRESENTATIVE

I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to participate in this study. I have been given a copy of this form.

_____

Name of Subject

_____

Name of Legal Representative (if applicable)

_____ _____

Signature of Subject or Legal Representative Date

SIGNATURE OF INVESTIGATOR

In my judgment the subject is voluntarily and knowingly giving informed consent and possesses the legal capacity to give informed consent to participate in this research study.

_____ _____

Signature of Investigator Date

E.2 Pre-experiment Survey

Demographic Survey

1. Subject number:_____

2. Age:_____

3. Gender: *M* *F*

4. Occupation:_____

   if student, (circle one): *Undergrad* *Masters* *PhD*

   expected year of graduation:_____

5. Military experience (circle one): *No* *Yes* If yes, which branch:_____

   Years of service:_____

6. How much sleep did you get for the past two nights?

   Last night:
   Night before last:

7. How often do you play computer games?

   Rarely *Monthly* *Weekly* *A few times a week* *Daily*

   Types of games played:_____

Boredom Proneness Scale (Farmer & Sundberg, 1986)

1.It is easy for me to concentrate on my activities.	T | F
2. Frequently when I am working I find myself worrying about other things.	T | F
3. Time always seems to be passing slowly.	T | F
4. I often find myself at "loose ends," not knowing what to do.	T | F
5. I am often trapped in situations where I have to do meaningless things.	T | F
6. Having to look at someone's home movies or travel slides bores me tremendously.	T | F
7. I have projects in mind all the time, things to do.	T | F
8. I find it easy to entertain myself.	T | F
9. Many things I have to do are repetitive and monotonous.	T | F
10. It takes more stimulation to get me going than most people.	T | F
11. I get a kick out of most things I do.	T | F
12. I am seldom excited about my work.	T | F
13. In any situation I can usually find something to do or see to keep me interested.	T | F
14. Much of the time I just sit around doing nothing.	T | F
15. I am good at waiting patiently.	T | F
16. I often find myself with nothing to do-time on my hands.	T | F
17. In situations where I have to wait, such as a line or queue, I get very restless.	T | F
18. I often wake up with a new idea.	T | F
19. It would be very hard for me to find a job that is exciting enough.	T | F
20. I would like more challenging things to do in life.	T | F
21. I feel that I am working below my abilities most of the time.	T | F
22. Many people would say that I am a creative or imaginative person.	T | F
23. I have so many interests, I don't have time to do everything.	T | F
24. Among my friends, I am the one who keeps doing something the longest.	T | F

25. Unless I am doing something exciting, even dangerous, I feel half-dead and dull.	T | F
26. It takes a lot of change and variety to keep me really happy.	T | F
27. It seems that the same things are on television or the movies all the time; it's getting old.	T | F
28. When I was young, I was often in monotonous or tiresome situations.	T | F

Personality Survey

NEO™-FFI-3

NEO Five-Factor Inventory-3

Paul T. Costa, Jr., PhD and Robert R. McCrae, PhD

Item Booklet | Form S–Adult

SELF-REPORT

Instructions

Write only where indicated in this Item Booklet. Carefully read all of the instructions before beginning. This questionnaire contains 60 statements. Read each statement carefully. For each statement, fill in the circle with the response that best represents your opinion. Make sure that your answer is in the correct box.

Fill in (SD) if you *strongly disagree* or the statement is definitely false.

Fill in (D) if you *disagree* or the statement is mostly false.

Fill in (N) if you are *neutral* on the statement, if you cannot decide, or if the statement is about equally true and false.

Fill in (A) if you *agree* or the statement is mostly true.

Fill in (SA) if you *strongly agree* or the statement is definitely true.

Note that the responses are numbered in ***rows***.

Example

First five responses from an individual who strongly disagrees with items 1, 2, and 3, and agrees with items 4 and 5.

Fill in only one response for each statement. Respond to all of the statements, making sure that you fill in the correct response. **DO NOT ERASE**! If you need to change an answer, make an "**X**" through the incorrect response and then fill in the correct response.

Before responding to the statements, turn to the inside of this Item Booklet and enter your name, age, sex, ID number (if any), and today's date.

PAR™ • 16204 N. Florida Ave. • Lutz, FL 33549 • 1.800.331.8378 • www.parinc.com

9 8 7 6 5 4 3 2 1 Reorder #RO-6806 Printed in the U.S.A.

ID# _____

Name _____ Age _____ Sex _____ Today's date _____

1. I am not a worrier.
2. I like to have a lot of people around me.
3. I enjoy concentrating on a fantasy or daydream and exploring all its possibilities, letting it grow and develop.
4. I try to be courteous to everyone I meet.
5. I keep my belongings neat and clean.
6. At times I have felt bitter and resentful.
7. I laugh easily.
8. I think it's interesting to learn and develop new hobbies.
9. At times I bully or flatter people into doing what I want them to.
10. I'm pretty good about pacing myself so as to get things done on time.
11. When I'm under a great deal of stress, sometimes I feel like I'm going to pieces.
12. I prefer jobs that let me work alone without being bothered by other people.
13. I am intrigued by the patterns I find in art and nature.
14. Some people think I'm selfish and egotistical.
15. I often come into situations without being fully prepared.
16. I rarely feel lonely or blue.
17. I really enjoy talking to people.
18. I believe letting students hear controversial speakers can only confuse and mislead them.
19. If someone starts a fight, I'm ready to fight back.
20. I try to perform all the tasks assigned to me conscientiously.
21. I often feel tense and jittery.
22. I like to be where the action is.
23. Poetry has little or no effect on me.
24. I'm better than most people, and I know it.
25. I have a clear set of goals and work toward them in an orderly fashion.
26. Sometimes I feel completely worthless.
27. I shy away from crowds of people.
28. I would have difficulty just letting my mind wander without control or guidance.
29. When I've been insulted, I just try to forgive and forget.
30. I waste a lot of time before settling down to work.
31. I rarely feel fearful or anxious.
32. I often feel as if I'm bursting with energy.
33. I seldom notice the moods or feelings that different environments produce.
34. I tend to assume the best about people.
35. I work hard to accomplish my goals.
36. I often get angry at the way people treat me.
37. I am a cheerful, high-spirited person.
38. I experience a wide range of emotions or feelings.
39. Some people think of me as cold and calculating.
40. When I make a commitment, I can always be counted on to follow through.

41. Too often, when things go wrong, I get discouraged and feel like giving up.
42. I don't get much pleasure from chatting with people.
43. Sometimes when I am reading poetry or looking at a work of art, I feel a chill or wave of excitement.
44. I have no sympathy for beggars.
45. Sometimes I'm not as dependable or reliable as I should be.

46. I am seldom sad or depressed.
47. My life is fast-paced.
48. I have little interest in speculating on the nature of the universe or the human condition.
49. I generally try to be thoughtful and considerate.
50. I am a productive person who always gets the job done.

51. I often feel helpless and want someone else to solve my problems.
52. I am a very active person.
53. I have a lot of intellectual curiosity.
54. If I don't like people, I let them know it.
55. I never seem to be able to get organized.

56. At times I have been so ashamed I just wanted to hide.
57. I would rather go my own way than be a leader of others.
58. I often enjoy playing with theories or abstract ideas.
59. If necessary, I am willing to manipulate people to get what I want.
60. I strive for excellence in everything I do.

Enter your responses here—remember to enter responses <u>*ACROSS*</u> the rows.

SD = *Strongly Disagree*; D = *Disagree*; N = *Neutral*; A = *Agree*; SA = *Strongly Agree*

ENTER ACROSS ⇨

1 SD D N A SA	2 SD D N A SA	3 SD D N A SA	4 SD D N A SA	5 SD D N A SA
6 SD D N A SA	7 SD D N A SA	8 SD D N A SA	9 SD D N A SA	10 SD D N A SA
11 SD D N A SA	12 SD D N A SA	13 SD D N A SA	14 SD D N A SA	15 SD D N A SA
16 SD D N A SA	17 SD D N A SA	18 SD D N A SA	19 SD D N A SA	20 SD D N A SA
21 SD D N A SA	22 SD D N A SA	23 SD D N A SA	24 SD D N A SA	25 SD D N A SA
26 SD D N A SA	27 SD D N A SA	28 SD D N A SA	29 SD D N A SA	30 SD D N A SA
31 SD D N A SA	32 SD D N A SA	33 SD D N A SA	34 SD D N A SA	35 SD D N A SA
36 SD D N A SA	37 SD D N A SA	38 SD D N A SA	39 SD D N A SA	40 SD D N A SA
41 SD D N A SA	42 SD D N A SA	43 SD D N A SA	44 SD D N A SA	45 SD D N A SA
46 SD D N A SA	47 SD D N A SA	48 SD D N A SA	49 SD D N A SA	50 SD D N A SA
51 SD D N A SA	52 SD D N A SA	53 SD D N A SA	54 SD D N A SA	55 SD D N A SA
56 SD D N A SA	57 SD D N A SA	58 SD D N A SA	59 SD D N A SA	60 SD D N A SA

Now answer the three questions labeled A, B, and C below.

A. Have you responded to all of the statements? _____ Yes _____ No

B. Have you entered your responses across the rows? _____ Yes _____ No

C. Have you responded accurately and honestly? _____ Yes _____ No

E.3 Post-experiment Survey

1. How confident were you about the actions you took?

   Not Confident Somewhat Confident Confident Very Confident Extremely Confident

2. How did you feel you performed?

   Very Poor Poor Satisfactory Good Excellent

3. How busy did you feel during the mission?

   Idle Not Busy Busy Very Busy Extremely Busy

4. Did you feel distracted at any point in the mission? *Yes No*

If so, please list some of the items or activities that distracted you from the mission:

5. Did you pay attention to aural alerts?

   Never Rarely Occasionally Frequently Always

6. How useful did you find aural alerts in improving your overall performance?

   Not useful Somewhat Useful Useful Very Useful Extremely Useful

7. Other comments:

Appendix F: Video Coding Criteria

1). Directed Attention

The participant appears focused and is only monitoring or interacting with the interface and not doing any other task.

2). Divided Attention

The participant has eyes on the interface screen, but multitasks in the following ways.
2p). Physiological diversions (examples: yawning, eating, fidgeting, stretching, and scratching)
2s). Social diversions (examples: talking, glancing at each other)
2c). Cognitive diversions (playing Minesweeper, flash games on the same screen as the simulation interface or glancing at the secondary monitor)

3). Distracted Attention

The participant is not paying attention to the interface at all.
3p). Physiological distractions (examples: sleeping, eating a meal without looking at the interface)
3s). Social distractions (examples: discussions with participants' backs turned to the computer)
3c). Cognitive distractions (reading a book, using the internet or other applications on the second screen, checking email and phone messages without looking back at interface)

Appendix G: Attention States, Subjective Metrics, and Personality Dimensions

G.1 Summary of Attention States

	1st session			2nd session			Without intervention			With intervention		
	Dir.	Div.	Dis.	Dir.	Div.	Dis.	Dir.	Div.	Dis.	Dir.	Div.	Dis.
Average (%)	69	22	9	58	27	15	62	27	11	66	22	12
SD (%)	9	4	6	8	5	6	7	4	4	9	6	6
Max (%)	88	35	20	89	44	31	88	44	20	82	35	31
Min (%)	56	9	2	37	4	4	37	9	2	40	4	2

Maximum and minimum values are based on max and min average values among participants
Dir. – Directed
Div. – Divided
Dis. – Distracted
SD – standard deviation

G.2 Percentages of Directed Attention State in 15 minute blocks

Time (min)	Without intervention (%)	With intervention (%)
0-15	80.2	85.8
15-30	72.1	84.1
30-45	64.2	72.4
45-60	65.6	63.2
60-75	56.6	66.4
75-90	69.6	68.6
90-105	65.1	68.4
105-120	56.4	59.9
120-135	51.3	55.1
135-150	58.1	61.8
150-165	53.7	60.7
165-180	62.7	53.8
180-195	59.0	57.6
195-210	55.9	66.6
210-225	66.9	64.1
225-240	59.9	59.7

G.3 Comparison of Percentages of Directed Attention State across the Two Scenarios

Paired Samples Test

		Paired Differences			t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean			
Pair 1	0-15	-5.61222	18.38506	6.12835	-.916	8	.387
Pair 2	15-30	-11.99669	17.33732	5.77911	-2.076	8	.072
Pair 3	30-45	-8.20188	29.19385	9.73128	-.843	8	.424
Pair 4	45-60	2.39807	32.96419	10.98806	.218	8	.833
Pair 5	60-75	-9.81104	19.94509	6.64836	-1.476	8	.178
Pair 6	75-90	.97673	27.29894	9.09965	.107	8	.917
Pair 7	90-105	-3.34828	20.27686	6.75895	-.495	8	.634
Pair 8	105-120	-3.48650	26.82746	8.94249	-.390	8	.707
Pair 9	120-135	-3.69473	18.91837	6.30612	-.586	8	.574
Pair 10	135-150	-3.80747	29.55629	9.85210	-.386	8	.709
Pair 11	150-165	-7.00088	28.71627	9.57209	-.731	8	.485
Pair 12	165-180	8.87039	24.07549	8.02516	1.105	8	.301
Pair 13	180-195	1.33347	18.57846	6.19282	.215	8	.835
Pair 14	195-210	-10.61942	31.31460	10.43820	-1.017	8	.339
Pair 15	210-225	2.85599	21.88186	7.29395	.392	8	.706
Pair 16	225-240	4.14953	27.16212	9.05404	.458	8	.659

Using Bonferroni's adjustment, a statistical difference between the percentages of directed attention states across the two scenarios can be established if $|t(8)|>4.169$. According to the above table, no statistical difference exists.

G.4: Descriptive Statistics of Subjective Metrics

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Confidence – without intervention	9	2.0	5.0	3.889	1.2693
Confidence – with intervention	9	2.0	5.0	3.556	1.2360
Attention to alerts without intervention	9	4.0	5.0	4.667	.5000
Attention to alerts - with intervention	9	4.0	5.0	4.778	.4410
Performance - without intervention	9	3.0	5.0	4.111	.7817
Performance - with intervention	9	2.0	5.0	4.000	1.0000
Busyness - without intervention	9	1.0	3.0	2.222	.6667
Busyness - with intervention	9	1.0	3.0	2.444	.7265
Usefulness of alerts - without intervention	9	1.0	5.0	3.778	1.3944
Usefulness of alerts - with intervention	9	2.0	5.0	3.667	1.0000
Valid N (listwise)	9				

G.5 Wilcoxon Signed Rank test of subjective data

Comparison between self-rated performance, busyness, confidence, attention to alerts and usefulness of alerts between the scenario without the design intervention and with the design intervention.

Test Statistics

	Performance self- rating	Busyness self-rating	Usefulness of alerts self-rating	Confidence self-rating	Attention to alerts self-rating
Z	-1.000	-.816	-.276	-1.732	-.447
p-value	.317	.414	.783	.083	.655

G.6 NEO Five Factor Inventory results on a scale of 25 to 75

Subject	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness
1	37	67	54	25	40
2	53	56	54	45	48
3	52	55	70	40	48
4	50	61	62	57	60
5	53	59	65	48	46
6	62	43	60	37	33
7	40	58	60	52	46
8	38	58	46	25	51
9	44	50	54	48	33

G.7 Descriptive Statistics of NEO Five Factor Inventory Survey

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Personality - Neuroticism	9	37.0	62.0	47.667	8.4113
Personality - Extraversion	9	43.0	67.0	56.333	6.7823
Personality - Openness	9	46.0	70.0	58.333	7.1414
Personality - Agreeableness	9	25.0	57.0	41.889	11.2522
Personality - Conscientiousness	9	33.0	60.0	45.000	8.6168
Valid N (listwise)	9				

G.8 Personality Dimension Comparisons with the Theoretical Mean

One-Sample Test

	Test Value = 50					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Personality - Neuroticism	-.832	8	.429	-2.3333	-8.799	4.132
Personality - Extraversion	2.801	8	.023	6.3333	1.120	11.547
Personality - Openness	3.501	8	.008	8.3333	2.844	13.823
Personality- Agreeableness	-2.163	8	.063	-8.1111	-16.760	.538
Personality - Conscientiousness	-1.741	8	.120	-5.0000	-11.623	1.623

G.9 Spearman's Personality Dimensions and Performance Score Correlations

Correlations

			Performance score (without intervention)	Performance score (with intervention)
Spearman's rho	Personality-Neuroticism	Correlation Coefficient	.276	-.418
		Sig. (2-tailed)	.472	.262
		N	9	9
	Personality-Extraversion	Correlation Coefficient	-.025	.243
		Sig. (2-tailed)	.949	.529
		N	9	9
	Personality-Openness	Correlation Coefficient	.247	.162
		Sig. (2-tailed)	.522	.678
		N	9	9
	Personality-Agreeableness	Correlation Coefficient	-.050	.168
		Sig. (2-tailed)	.897	.666
		N	9	9
	Personality-Conscientiousness	Correlation Coefficient	.641	.489
		Sig. (2-tailed)	.063	.181
		N	9	9

G.10 Spearman's Personality Dimensions and Directed Attention State Correlations

Correlations

			Directed –with intervention	Directed –without intervention
Spearman's rho	Personality - Neuroticism	Correlation Coefficient	-.243	-.611
		Sig. (2-tailed)	.529	.081
		N	9	9
	Personality - Extraversion	Correlation Coefficient	-.209	-.033
		Sig. (2-tailed)	.589	.932
		N	9	9
	Personality - Openness	Correlation Coefficient	-.417	-.264
		Sig. (2-tailed)	.264	.493
		N	9	9
	Personality - Agreeableness	Correlation Coefficient	.269	.160
		Sig. (2-tailed)	.484	.682
		N	9	9
	Personality - Conscientiousness	Correlation Coefficient	.177	.118
		Sig. (2-tailed)	.648	.762
		N	9	9

Appendix H: BPS Scores, Correlations, and Utilization

H.1 Utilization & Boredom Proneness Scale (BPS) Score on a scale of 0 to 28

Subject #	BPS Score	Utilization - without intervention (%)	Utilization - with intervention (%)
1	10	12.6	10
2	8	22.9	22.3
3	7	10.4	9.5
4	4	21.1	18.6
5	4	8.3	11.4
6	11	15.7	17
7	4	13.7	16.7
8	16	11.9	12.8
9	6	13.6	11.5

H.2 Utilization Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Utilization – without intervention	9	8	23	14.47	4.778
Utilization – with intervention	9	10	22	14.42	4.408
Valid N (listwise)	9				

H.2 Pearson's BPS score and performance score correlations

Correlations

		BPS Score
Performance score (without intervention)	Pearson Correlation	.346
	Sig. (2-tailed)	.362
	N	9
Performance score (with intervention)	Pearson Correlation	.188
	Sig. (2-tailed)	.629
	N	9

H.3 Spearman's BPS score and directed attention state correlations

Correlations

			Boredom Proneness
Spearman's rho	Directed attention – with intervention	Correlation Coefficient	-.017
		Sig. (2-tailed)	.965
		N	9
	Directed attention- without intervention	Correlation Coefficient	-.068
		Sig. (2-tailed)	.862
		N	9

References

Aaronson, L. S., Teel, C. S., Cassmeyer, V., Neuberger, G. B., Pallikkathayil, L., Pierce, J., et al. (1999). Defining and Measuring Fatigue. *Nursing Research, 31*(1), 45-50.

Achermann, P. (2004). The Two-Process Model of Sleep Regulation Revisited. *Aviation, Space, and Environmental Medicine, 75*(3, Suppl.), A37-43.

Bainbridge, L. (1983). Ironies of Automation. *Automatica, 19*, 775-779.

Balci, O. (1997). *Verification, Validation and Accreditation of Simulation Models.* Paper presented at the 1997 Winter Simulation Conference, Atlanta, GA.

Balci, O. (2003). *Verification, Validation, and Certification of Modeling and Simulation Applications.* Paper presented at the 2003 Winter Simulation Conference, New Orleans, LA.

Balsamo, S., Persone, V., & Onvural, R. (2001). *Analysis of Queuing Networks with Blocking*. Boston, MA: Kluwer Academic Publisher.

Banks, J., Carson, J. S., Nelson, B. L., & Nicol, D. M. (2005). *Discrete-Event System Simulation* (4th ed.): Pearson Education.

Becker, A. B., Warm, J. S., Dember, W. N., & Hancock, P. A. (1991). *Effects of Feedback on Perceived Workload in Vigilance Performance.* Paper presented at the Human Factors and Ergonomics Society 35th Annual Meeting, San Francisco, CA.

Belton, T., & Priyadharshini, E. (2007). Boredom and Schooling: a Cross-Disciplinary Exploration. *Cambridge Journal of Education, 37*(4), 579-595.

Belyavin, A. J., & Spencer, M. B. (2004). Modeling Performance and Alertness: The QinetiQ Approach. *Aviat Space Environ Med, 75*(3, Suppl.), A93-103.

Berger, P. J., McCutcheon, L., Soust, M., Walker, A. M., & Wilkinson, M. H. (1991). Electromyographic Changes in the Isolated Rat Diaphragm During the Development of Fatigue. *European Journal of Applied Physiology, 62*, 310-316.

Borbely, A., & Achermann, P. (1999). Sleep Homeostasis and Models of Sleep Regulation. *Biological Rhythms, 14*(6), 557-568.

Bortscheller, B. J., & Saulnier, E. T. (1992). *Model Reusability in a Graphical Simulation Package.* Paper presented at the Winter Simulation Conference, Arlington, VA.

Brainbridge, L. (1987). Ironies of Automation. In J. Rasmussen, K. Duncan & J. Leplat (Eds.), *New Technology and Human Error*. New York: Wiley.

Britton, A., & Shipley, M. J. (2010). Bored to death? *International Journal of Epidemiology, 39*(2), 370-371.

Brodsky, J. (1995). In Praise of Boredom *On Grief and Reason*. New York: Farrar Strauss Publishing.

Brown, G. H., & Carroll, C. D. (1984). *The Effect of Anxiety and Boredom on Cognitive Test Performance.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Button, K. (2009). Different Courses: New Style UAV Trainees Edge Toward Combat. *Training and Simulation Journal*, 42-44.

Carbonell, J. R. (1966). A Queuing Model of Many-Instrument Visual Sampling. *IEEE Transactions on Human Factors in Electronics, 4*(4), 157-164.

Carpenito, L. J. (1995). *Nursing Diagnosis: Application to Clinical Practice*. Philadelphia: J.B. Lippincott.

Chanel, G., Rebetez, C., Betrancourt, M., & Pun, T. (2008). *Boredom, Engagement and Anxiety as Indicators for Adaption to Difficulty in Games.* Paper presented at the 12th International Conference on Entertainment and Media in the Ubiquitous Era, New York, NY.

Chu, Y., & Rouse, W. B. (1979). Adaptive Allocation of Decisionmaking Responsibility Between Human and Computer in Multitask Situations. *IEEE Transactions on Systems, Man, and Cybernetics, 9*(12), 769-778.

Colligan, M. J., & Murphy, L. R. (1979). Mass Psychogenic Illness in Organizations: An overview. *Journal of Occupational Psychology, 52*(2), 77-90.

Costa, P. T., & McCrae, R. R. (1992). *Neo-Personality Inventory - Revised (NEO PI-R)*: Psychological Assessment Resources.

Cummings, M., & Nehme, C. (2009). *Modeling the Impact of Workload in Network Centric Supervisory.* Paper presented at the 2nd Annual Sustaining Performance Under Stress Symposium, College Park, MD.

Cummings, M. L. (2008). Of Shadows and white scarves. *C4ISR Journal, August*.

Cummings, M. L., & Guerlain, S. (2007). Developing Operator Capacity Estimates for Supervisory Control of Autonomous Vehicles. *Human Factors, 49*(1), 1-15.

D'Mello, S., Chapman, P., & Graesser, A. (2007). *Posture as a Predictor of Learner's Affective Engagement.* Paper presented at the 29th Annual Meeting of the Cognitive Science Society, Austin, TX.

Davies, D. R., & Parasuraman, R. (1982). *The Psychology of Vigilance*. London: Academic Press.

Dittmar, M. L., Warm, J. S., Dember, W. N., & Ricks, D. F. (1993). Sex Differences in Vigilance Performance and Perceived Workload. *The Journal of General Psychology, 120*(3), 309-322.

Donmez, B., Nehme, C., & Cummings, M. L. (2010). Modeling Workload Impact in Multiple Unmanned Vehicle Supervisory Control. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 40*(6), 1180-1190.

Dostoevsky, F. (1997). *A Writer's Diary* (Vol. II). Evanston, IL: Northwestern University Press.

Ebrahim, S. (2010). Rapid Responses, Population Prevention and Being Bored to Death. *International Journal of Epidemiology, 39*(2), 323-326.

Farmer, R., & Sundberg, N. D. (1986). Boredom Proneness - The Development and Correlates of a New Scale. *Journal of Personality Assessment, 50*(1), 4-17.

Fisher, C. D. (1993). Boredom at Work: A Neglected Concept. *Human Relations, 46*, 395-417.

Fisher, S. (2008). *Replan Understanding for Heterogenous Unmanned Vehicle Teams.* Massachusetts Institute of Technology, Cambridge, MA.

Grubb, E. A. (1975). Assembly Line Boredom and Individual Differences in Recreational Participation. *Leisure Research, 7*, 256-269.

Haga, S. (1984). An Experimental Study of Signal Vigilance Errors in Train Driving. *Ergonomics, 27*, 755-765.

Hart, C. S. (2010). *Assessing the Impact of Low Workload in Supervisory Control of Networked Unmanned Vehicles.* Massachusetts Institute of Technology, Cambridge, MA.

Hitchcock, E. M., Dember, W. N., Warm, J. S., Moroney, B. W., & See, J. E. (1999). Effects of Cueing and Knowledge of Results on Workload and Boredom in Sustained Attention. *Human Factors, 41*(3), 365-372.

Hursh, S. R., Redmond, D. P., Johnson, M. L., Thorne, D. R., Belenky, G., Balkin, T. J., et al. (2004). Fatigue models for applied research in warfighting. *Aviation, Space, and Environonmental Medicine, 75*(3, Suppl.), A44-53.

Iso-Ahola, S. E., & Weissinger, E. (1987). Leisure and Boredom. *Social and Clinical Psychology, 5*, 356-364.

Jacobs, M., Fransen, B., McCurry, J., Heckel, F. W. P., Wagner, A. R., & Trafton, J. (2009). *A Preliminary System for Recognizing Boredom.* Paper presented at the ACM/IEEE International Conference on Human-Robot Interaction, La Jolla, CA.

Jewett, M. E., & Kronauer, R. E. (1999). Interactive Mathematical Models of Subjective Alertness and Cognitive Throughput in Humans. *Biological Rhythms, 14*(6), 588-597.

Jung, T. P., Makeig, S., Stensmo, M., & Sejnowski, T. J. (1997). Estimating Alertness from the EEG Power Spectrum. *IEEE Transactions on Biomedical Engineering, 44*(1), 60-60.

Kaku, M., & Trainer, J. (1992). *Nuclear Power: Both Sides: The Best Arguments For and Against the Most Controversial Technology*. New York: W. W. Norton & Company, Inc.

Kass, S. J., Vodanovich, S. J., Stanny, C., & Taylor, T. M. (2001). Watching the Clock: Boredom and Vigilance Performance. *Perceptual Motor Skills, 92*, 969-976.

Kendall, D. G. (1953). Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of Embedded Markov Chains. *Annals of Mathematical Statistics, 24*, 338-354.

Krupp, L. B., LaRocca, N. G., Muir-Nash, J., & Steinberg, A. D. (1989). The Fatigue Severity Scale. Application to Patients with Multiple Sclerosis and Systematic Lupus Erythematosus. *Archives of Neurology, 46*, 1121-1123.

Lan, P., Ji, Q., & Looney, C. (2003). *Non-intrusive real time human fatigue modeling and monitoring.* Paper presented at the Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society, Denver, CO.

Langan-Fox, J., Sankey, M. J., & Canty, J. M. (2008). Keeping the Human in the Loop: From ATCOs to ATMs, *Keynote Speech by J. Langan-Fox at the Smart Decision Making for Clean Skies Conference*. Canberra, Australia.

Langan-Fox, J., Sankey, M. J., & Canty, J. M. (2009). Human Factors Measurement for Future Air Traffic Control Systems. *Human Factors, 51*, 595-637.

Lee, K. A., Hicks, G., & Nino-Murcia, G. (1991). Validity and Reliability of a Scale to Assess Fatigue. *Psychiatry Research, 36*, 291-298.

Liu, Y., Feyen, R., & Tsimhoni, O. (2006). Queuing Network-Model Human Processor (QN-MHP): A Computational Architecture for Multitask Performance in Human-Machine Systems. *ACM Transactions on Computer-Human Interaction, 13*(1), 39-46.

Mackworth, N. H. (1950). *Researches on the Measurement of Human Performance (Special Report No 268)*. London: Medical Reseatch Council, Her Majesty's Stationary Office

Mackworth, N. H. (1957). Some Factors Affecting Vigilance. *Advancement of Science, 53*, 389-393.

Mallis, M. M., Mejdal, S., Nguyen, T. T., & Dinges, D. F. (2004). Summary of the Key Features of Seven Biomathematical Models of Human Fatigue and Performance. *Aviation, Space, and Environmental Medicine, 75*(1), A4-14.

Manly, T., Robertson, I. H., Galloway, M., & Hawkins, K. (1999). The Absent Mind: Further Investigations of Sustained Attention to Response. *Neuropsychologia, 37*, 661-670.

McCorkle, R., & Young, K. (1978). Develpment of a Symptom Distress Scale. *Cancer Nursing, 1*, 373-378.

McNair, D. M., Lorr, M., & Droppleman, L. F. (1992). *EdITS Manual for the Profile of Mood States*. San Diego, CA: EdITS/Educational and Industrial Testing Service.

McRae, R. R., & Costa, P. T. (2010). *NEO™ Inventories*. Lutz, FL: PAR Publishing.

Merriam-Webster Dictionary. (2011). Boredom - Definition and More. Retrieved April 11, 2011, from http://www.merriam-webster.com/dictionary/boredom

Moore-Ede, M., Heitmann, A., Guttkuhn, R., Trutschel, U., Aguirre, A., & Croke, D. (2004). Circadian Alertness Simulator for Fatigue Risk Assessment in Transportation: Application to Reduce Frequency and Severity of Truck Accidents. *Aviat Space Environ Med, 75*(3, Suppl.), A119-121.

Muscio, B. (1921). Is a Fatigue Test Possible? *British Journal of Psychology, 12*, 31-46.

Naylor, T. H., & Finger, J. M. (1967). Verification of Computer Simulation Models. *Management Science, 2*, B92-B101.

Nehme, C. (2009). *Modeling Human Supervisory Control in Heterogeneous Unmanned Vehicle Systems.* Massachusetts Institute of Technology, Cambridge, MA.

Nehme, C., Crandall, J., & Cummings, M. (2008). *Using Discrete Event Simulation to Model Situational Awareness of Unmanned-Vehicle Operators.* Paper presented at the Modeling, Analysis and Simulation Center Capstone Conference, Norfolk, VA.

Nehme, C., Kilgore, R. M., & Cummings, M. L. (2008, September 22-26). *Predicting the Impact of Heterogeneity on Unmanned-Vehicle Team Performance.* Paper presented at the 52nd Annual Meeting of the Human Factors and Ergonomic Society, New York, NY, USA.

Nehme, C., Mekdeci, B., Crandall, J. W., & Cummings, M. L. (2008). The impact of heterogeneity on operator performance in futuristic unmanned vehicle systems. *The International C2 Journal, 2*(2), 1-30.

Parasuraman, R., & Davies, D. R. (1976). Decision Theory Analysisof Response Latencies in Vigilance. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 569-582.

Pattyn, N., Neyt, X., Henderickx, D., & Soetens, E. (2008). Psychological Investigation of Vigilance Decrement: Boredom or Cognitive Fatigue? *Physiology & Behavior, 93*(1-2), 369-378.

Pina, P. E., Cummings, M. L., Crandall, J. W., & Della Penna, M. (2008). Identifying Generalizable Metric Classes to Evaluate Human-Robot Teams *Proceedings of Metrics for Human-Robot Interaction Workshop at the 3rd Annual Conference on Human-Robot Interaction*. Amsterdam, The Netherlands.

Pinedo, M. (2002). *Scheduling: Theory, Algorithms, and Systems* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Piper, B. F. (1989). Fatigue: Current Bases for Practice. In S. G. Funk, E. M. Tornquist, M. T. Champagne & R. Wiese (Eds.), *Key Aspects of Comfort*. New York: Springer.

Posner, J., Russell, J. A., Gerber, A., Gorman, D., Colibazzi, T., Yu, S., et al. (2008). The Neurophysiological Bases of Emotion: An fMRI Study of the Affective Circumplex using Emotion-denoting Words. *Human Brain Mapping, 30*(3), 883-895.

Prinzel III, L. J., DeVries, H., Freeman, F. G., & Mikulka, P. (2001). Examination of Automation-Induced Complacency and Individual Difference Variates, *NASA/TM-2001-211413*. Langley Research Center: National Aeronautics and Space Administration.

Prinzel III, L. J., & Freeman, F. G. (1997). Task-specific Sex Differences in Vigilance Performance: Subjective Workload and Boredom. *Perceptual and Motor Skills, 85*(3), 1195-1202.

Proctor, R. W., & Zandt, T. V. (2008). *Human Factors in Simple and Complex Systems* (2nd ed.). Boca Raton, FL: CRC Press.

Ragheb, M. G., & Merydith, S. P. (2001). Development and Validation of a Unidimensional Scale Measuring Free Time Boredom. *Leisure Studies, 20*, 41-59.

Rhoten, D. (1982). Fatigue and the Postsurgical Patient. In C. M. Norris (Ed.), *Concept Clarification in Nursing* (pp. 277-300). Rockville, MD: Aspen.

Rifkin, J. (1995). *The End of Work*. New York: Putman's Publishing.

Rodgers, M. D., & Nye, L. G. (1993). Factors Associated with Severity of Operational Errors at Air Route Traffic Control Centers. In M. D. Rodgers (Ed.), *An Examination of the Operational Error Database for Air Traffic Control Centers* (pp. 243-256). Washington D.C.: Federal Aviation Administration, Office of Aviation Medicine.

Sargent, R. (2007). *Verification and Validation of Simulation Models.* Paper presented at the 2007 Winter Simulation Conference, Washington, DC.

Sawin, D. A., & Scerbo, M. W. (1994). *Vigilance: How to Do it and Who Should Do It.* Paper presented at the The Human Factors and Ergonomics Society 38th Annual Meeting, Nashville, TN.

Sawin, D. A., & Scerbo, M. W. (1995). Effects of Instruction Type and Boredom Proneness in Vigilance: Implications for Boredom and Workload. *Human Factors, 37*(4), 752-765.

Scerbo, M. W. (1998). What's So Boring About Vigilance? In R. R. Hoffman, M. F. Sherrick & J. S. Warm (Eds.), *Viewing Psychology as a Whole: The Integrative Science of William N. Dember* (pp. 145-166): American Psychological Association.

Schmidt, D. K. (1978). A Queuing Analysis of the Air Traffic Controller's Workload. *IEEE Transactions on Systems, Man, and Cybernetics, 8*(6), 492-498.

Schmidt, E. A., Kincses, W. E., Schrauf, M., Haufe, S., Schubert, R., & Curio, G. (2006). *Assesing Srivers' Vigilance State During Monotonous Driving.* Paper presented at the Proceedings of the Fourth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Stevenson, WA.

Schroeder, D. J., Touchstone, R. M., Stern, N., Stoliarov, N., & Thackray, R. I. (1994). Maintaining Vigilance on a Simulated ATC Monitoring Task Across Repeated Sessions. Oklahoma City: Federal Aviation Administration, Civil Aeromedical Institute.

Shaw, T. H., Warm, J. S., Finomore, L., Tripp, G., Matthews, E. W., & Parasuraman, R. (2009). Effects of Sensory Modality on Cerebral Blood Flow Velocity During Vigilance. *Neuroscience Letters, 461*(3), 207-211.

Sheridan, T. (1992). *Telerobotics, Automation and Human Supervisory Control*. Cambridge, MA: The MIT Press.

Snyder-Halpern, R., & Verran, J. A. (1987). Instrumentation to Describe Subjective Sleep Characteristics in Healthy Subjects. *Research in Nursing and Health, 10*, 155-163.

Spartacus Educational. (2011). The Luddites. Retrieved February 13, 2011, from http://www.spartacus.schoolnet.co.uk/PRluddites.htm

Stager, P., Hameluck, D., & Jubis, R. (1989). *Underlying Factors in Air Traffic Control Incidents.* Paper presented at the Human Factors Society 33rd Annual Meeting, Denver, CO.

Stollery, B. T. (2006). Vigilance. In W. Karwowski (Ed.), *International Encyclopedia of Ergonomics and Human Factors* (Vol. 1, pp. 965-968). Boca Raton, FL: CRC Press.

Struss, D. T., Shallice, T., Alexander, M. P., & Picton, T. W. (1995). A Multidisciplinary Approach to Anterior Attentional Functions. *Annals of the NY Academy of Sciences, 769*(1), 191-212.

Svendsen, L. F. H. (2005). *A philosophy of boredom* (J. Irons, Trans.). London: Reaktion Books Ltd.

Tack, B. (1991). Dimensions and Correlates of Fatigue in Older Adults with Rheumtoid Arthritis. Unpublished Doctoral Dissertation. University of California.

Thackray, R. I. (1980). Boredom and Monotony as a Consequence of Automation: A Consideration of the Evidence Relating Boredom and Monotony to Stress, *DOT/FAA/AM-80/1*. Oklahoma City: Federal Aviation Administration, Civil Aeromedical Institute.

Thackray, R. I., Powell, J., Bailey, M. S., & Touchstone, R. M. (1975). Physiological, Subjective, and Performance Correlates of Reported Boredom and Monotony While Performing a Simulated Radar Control Task. Oklahoma City: Federal Aviation Administration, Civil Aeromedical Institute.

Thackray, R. I., & Touchstone, R. M. (1988). An Evaluation of the Effects of High Visual Taskload on the Separate Behaviors Involved in Complex Monitoring Performance. Oklahoma City: Federal Aviation Administration, Civil Aeromedical Institute.

The New York Times. (2009). Report on Pilots Who Overshot Airport. Retrieved February 10, 2011, from http://www.nytimes.com/2009/12/17/us/17pilot.html

Thompson, W. T., Lopez, N., Hickey, P., DaLuz, C., & Caldwell, J. L. (2006). *Effects of Shift Work and Sustained Operations: Operator Performance in Remotely Piloted Aircraft (OP-REPAIR)*. Paper presented at the 311th Performance Enhancement Directorate and Performance Enhancement Research Division.

Valenti, M., Bethke, B., Fiore, G., How, J. P., & Feron, E. (2006). *Indoor Multi-Vehicle Flight Testbed for Fault Detection, Isolation, and Recovery.* Paper presented at the AIAA Guidance, Navigation, and Control Conference, Keystone, CO.

Walden, R. S., & Rouse, W. B. (1978). A Queuing Model of Pilot Decisionmaking in a Multitask Flight Management System. *IEEE Transactions on Systems, Man, and Cybernetics, 8*(12), 867-875.

Warm, J. S., Dember, W. N., & Hancock, P. A. (1996). Vigilance and Workload in Automated Systems *Automation and Human Performance* (pp. 183-200). Mahwah, NJ: Erlbaum Publishing.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance Requires Hard Mental Work and is Stressful. *Human Factors, 50*(3), 433-441.

Weinger, M. B. (1999). Vigilance, Boredom, and Sleepiness. *Journal of Clinical Monitoring and Computing, 15*(7-8), 549-552.

Wellbrink, J., Zyda, M., & Hiles, J. (2004). Modeling Vigilance Performance as a Complex Adaptive System. *JDMS, 1*(1), 29-42.

Wickens, C. D. (2008). Multiple Resources and Mental Workload. *Human Factors, 50*(3), 449-455.

Wickens, C. D., & Hollands, J. G. (2000). *Engineering Psychology and Human Performance* (3rd ed.). Upper Saddle River, N.J.: Prentice Hall.

Winter, R. (2002). *Still Bored in a Culture of Entertainment: Rediscovering Passion & Wonder*. Nottingham, UK: IVP Books Publishing.

Zuckerman, M. (1979). *Sensation Seeking: Beyong the Optimal Level of Arousal*. Hillside, NJ: Erlbaum Publishing.